THE UNIVERSITY
of EDINBURGH

ACROSS BREED GENOMIC EVALUATIONS IN CATTLE

THE UNIVERSITY *of* EDINBURGH

Alexandra Brown

Thesis submitted for the degree of Doctor of Philosophy

Royal (Dick) School of Veterinary Studies
University of Edinburgh

2016

Declaration

I declare that this thesis is my own composition and that the research described in it is my own work, except where acknowledged. The work described has not been submitted for any other degree or professional qualification.

Alexandra Brown

29th September 2016

Abstract

Genomic evaluation techniques have been a huge success in the dairy cattle industry, as they allow accurate enough estimation of breeding values at a young age to allow selection decisions to be made at an earlier stage, thereby increasing the rate of genetic progress per annum. The success of genomic selection techniques relies on the existence of linkage disequilibrium (LD) between markers and quantitative trait loci (QTL) across the population of interest; LD persists across larger distances within breeds than across breeds. Therefore, most success so far has been for selection within breeds, but the industry is keen for "across breed" evaluations to be developed, both in a multi-breed scenario which would allow evaluations for breeds that are numerically too small to carry out evaluations within breeds, and also for the evaluation of crossbred animals.

This thesis investigates the potential for applying genomic selection techniques in both the multi-breed and crossbred scenarios. Chapter 2 examines the potential for a multi-breed reference population to improve the accuracy of genomic evaluation for a numerically small breed, for a range of production and non-production traits. The results provide evidence that forming a multi-breed reference population for two closely related breeds (Holstein and Friesian) results in a higher accuracy of GEBVs for the smaller breed, particularly when more phenotypic records are added via the single-step GBLUP method, and when a higher density SNP chip is used. Chapter 3 examines the crossbred scenario, whereby GEBVs are calculated for crossbred individuals based on a crossbred reference population. The population used for analysis was a highly crossbred African population, and GEBVs were calculated for

three groups of animals chosen according to whether they had a high or low proportion of imported dairy genetics. Accuracy of prediction was higher than expected, and provided proof of concept for applying genomic selection techniques in crossbred African cattle populations. Chapter 4 investigates the potential for using novel SNPs derived from sequence data in order to estimate genomic relationships across cattle breeds, deploying data from two closely related breeds, Fleckvieh and Simmental, and a further distant European breed, the Brown Swiss. Novel SNPs were selected from sequence based on their putative impact on the genome, with impacts being inferred by SNP annotation software snpEff. Results showed that genomic relationships calculated using novel SNPs have a high correlation with genomic relationships calculated using SNPs common to the Illumina BovineHD SNP chip, though between-breed correlations were lower than those within breeds.

The results presented in this thesis demonstrate that utilising a multi-breed reference population can improve the accuracy of prediction for a numerically small breed, and that genomic prediction of highly crossbred individuals is also feasible. However, differences between breeds and also types of crossbred animal suggest that no one solution can be used for all across-breed evaluations, and further research will be needed to allow commercial implementation in further populations.

Lay summary

Genomic selection allows us to predict the genetic merit of individual animals using a combination of information on the animal's measurable characteristics, such as milk yield or fertility, and pieces of its DNA, which are commonly called genetic markers. It offers an advantage to breeders over traditional selection techniques as it allows them to make accurate selection decisions at an earlier age, which leads to an increase in the rate of genetic progress, and therefore a better financial return in a fixed period of time. Genomic selection is currently applied commercially within individual cattle breeds, but research is on-going to successfully adapt the technique for use in populations made up of multiple breeds and also in populations of crossbred animals.

This research project looked at three scenarios for carrying out genomic selection across breeds. The first study looked at combining data from two closely-related breeds to improve the accuracy of prediction, and showed that including data from a larger population of closely-related animals does improve prediction accuracy. The second study explored genomic selection in an African crossbred cattle population where traditional selection techniques are not implemented, and showed that genomic prediction could be used to increase the rate of genetic progress in this population. The third study looked at whether it was possible to select a subset of genetic markers for genomic selection from the animals entire DNA sequence, as opposed to using the standard set of markers used commercially for genomic selection. Results from the selected subset of markers correlated highly with results from the standard panel of markers.

Results from this research project suggest that across-breed genomic evaluations are feasible, but should be further tested using larger datasets before commercialisation.

Publications

A Brown, J Ojango, J Gibson, M Coffey, M Okeyo, R Mrode. Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics Project for East Africa. *J Dairy Sci* 2016; **99**(9): 7308–7312

Conference contributions

A Brown, J Ojango, M Okeyo, R Mrode. Investigating genomic selection in crossbred cattle using data from the Dairy Genetics Project for East Africa. British Society of Animal Science Annual Conference. Chester, UK, 2016

A Brown, G Banos, M Coffey, J Woolliams, R Mrode. Estimation of genomic breeding values in a multi-breed context. British Society of Animal Science Annual Conference. Chester, UK, 2015

A Brown, G Banos, MP Coffey, JA Woolliams, RA Mrode. Holstein-Friesian relationships and the impact on accuracy of an across-breed evaluation. 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada, 2014

Acknowledgements

Firstly, I would like to thank the BBSRC, the Knowledge Transfer Network and AHDB Dairy, who provided the funds for me to carry out this work.

Huge thanks are due to my supervisors, Raphael Mrode, Georgios Banos, Mike Coffey and John Woolliams, who were generous enough to employ me for this project, and have been a pleasure to work with. Raph, thank you for your patience, for taking the time to share with me your vast knowledge, and for always greeting me with a big smile, even in times of despair! I couldn't have asked for a nicer primary supervisor. To Georgios, I am so grateful that you welcomed me into your wonderful team, thank you for being there for the day to day when Raph was away. To Mike, thank you for always having time to talk things through with me in my moments of panic, and your "morale boosting" chats. To John, an inspiring teacher, thank you for believing in my ability even when I didn't, and for taking the time to ask me all the awkward questions, and then patiently explain the answers when I quietly admitted that I didn't know!

To Marco Winters and Fern Pearston at AHDB Dairy, thank you for letting me join you for three really enjoyable months at AHDB, I look forward to working together again soon.

Thanks to colleagues from ILRI, particularly Julie Ojango and Ally Okeyo Mwai, for allowing me to work on the Dairy Genetics East Africa project data that comprises chapter three.

There are a number of colleagues within SRUC and Roslin without whose help I would never have survived the past four years, I'd like to thank Enrique Sánchez Molano and Joanna Ilska-Warner for always being willing to discuss my work and help me wade through the mess in my brain, Janez Jenko for all his help and advice regarding the sequence data, and everyone in EGENES, particularly Neil Clelland and John Stainton for their help with the Friesian genotypes, and Tomasz

Krzyzelewski and Umer Tahir for their help in accessing and sorting through data, and extracting huge pedigrees. Tomasz is a font of computing knowledge and taught me a lot, in return I was able to teach him that telling new students unfamiliar with UNIX to "just type delete *", probably isn't a good idea. Thanks to Arjan Tolkamp who did a stellar job as my go to programming guy, and most of all, a massive thank you to Sebastian Mucha, who put up with four whole years of silly and sometimes not so silly questions, helped me move from being an R hater to an R lover (ish!), and kept his promise not to move on to pastures new until I had submitted – good luck in your new role, they are so lucky to have you.

Thanks to the whole Banos team who welcomed me into the fold two years ago, I was so happy to be given an environment where I could present and discuss my results with no fear. You gave me valuable advice for all my chapters, and I hope I was able to return the favour.

Thank you to my friends who made my time in Edinburgh so enjoyable. To the "BaB" gang, thank you for the slightly weird but always fascinating coffee break and Whatsapp chats, and for the support during the low moments. Special mentions to Suzanne Desire who frequently dealt with my appearance at her desk with a face of doom, and to Jo Pollock, who has been my official "thesis buddy" since we shared our first day together. Thanks also to Dena Richarz for taking on the role of awesome flatmate; I'm sorry I restricted you to only one month of Christmas decorations per year!

If you had asked the 20-year old me – an undergraduate with below average grades and lacking in motivation – the typical "Where do you see yourself in ten years?" interview question, though I'm not sure exactly what I'd have answered, it definitely wouldn't have been "completing a PhD thesis". However, just before my final year of university, I was lucky to have the opportunity to undertake a summer placement with Dr Sarah Blott, completing a small project on parentage testing in Exmoor ponies. The chance to apply my effort to something that actually fixed a real-world

problem re-ignited my passion for science. Sarah, thank you for being a wonderful mentor, for believing in my potential and helping me in the first stages of my career, I absolutely wouldn't have got this far without you.

Thank you to my family for their support, I know they will also be relieved that it is nearly over! Also an unexpected thanks to my parents for sending me to participate in poetry and literature festivals when I was younger, Despite me being distinctly unimpressed at the time, it turns out that the experience prepared me well to be able to stand up in public and present my work.

Finally, I would like to thank James, who backed me when I swanned off to Scotland for a year to do a Masters, and didn't complain when one year away turned into five! (On that note, thanks to Easyjet and Ryanair for offering flights cheaply enough to enable my frequent trips home.) Thank you for all your love and support, and for tirelessly renovating our house while I was away – I'm eagerly looking forward to our PhD-free (and DIY-free!) future together!

List of abbreviations

BLUP	Best Linear Unbiased Predictor
EBV	Estimated Breeding Value
GEBV	Genomic Estimated Breeding Value
GBLUP	Genomic Best Linear Unbiased Predictor
HBLUP	Single-step Genomic Best Linear Unbiased Predictor
HD	High Density
SNP	Single Nucleotide Polymorphism
LD	Linkage Disequilibrium
Ne	Effective population size
QTL	Quantitative Trait Locus
MCMC	Markov Chain Monte Carlo
MAS	Marker Assisted Selection
PCA	Principal Components Analysis
YD	Yield Deviation
dEBV	De-regressed Estimated Breeding Value
SCC	Somatic Cell Count
WGS	Whole-Genome Sequence
BSW	Brown Swiss
FLE	Fleckvieh
SIM	Simmental
VR1	VanRaden's 1^{st} G
VR2	VanRaden's 2^{nd} G
VCF	Variant Call Format

List of tables

List of figures

Table of contents

Chapter 1: General Introduction

1.1 Traditional methods of genetic improvement

For thousands of years, humans have been making selection decisions for the improvement of livestock by using the most productive animals for breeding purposes (Hill, 2014), with advances in selective breeding being pioneered in the 18th century by Robert Bakewell (Orel, 1997). Genetic progress was evident but gradual until the mid-20th century, at which point a considerable increase in the rate of genetic improvement in both livestock and plant populations was observed. A good illustration of this is the increase in average milk yield of US Holstein cattle, which increased from around 6,000kg per lactation in 1960, to almost 12,000kg per lactation in 2000, with about 50% of this increase being credited to improved genetics (Dekkers and Hospital, 2002). This large improvement was made possible using quantitative genetics techniques, which use mathematical modelling to estimate the genetic and environmental components of the trait phenotype. This quantitative genetic approach is generally known as a black box approach, because the genetic architecture of the trait of interest is unknown. The underlying assumption of this approach is that an "infinite" number of genes each have a very small effect on the trait of interest, and this is known as the infinitesimal model (Dekkers and Hospital, 2002; Hill, 2010). In 1963 Henderson proposed a method which later came to be known as Best Linear Unbiased Prediction (BLUP) (Henderson, 1975). This method incorporates phenotypic and pedigree information into a linear mixed model to estimate individual breeding values (EBVs) which are used as an indication of the additive genetic merit of individuals. This technique swiftly became the method of choice for evaluating the genetic merit of potential breeding stock.

BLUP EBVs have long been the method of choice for selective breeding programmes in the dairy industry. EBVs were originally calculated for production traits (milk yield, protein content, fat content), but as breeding goals have broadened, the dairy industry now routinely calculates EBVs for a range of other traits, including type traits, fertility, mastitis, bovine tuberculosis resistance, lifespan and calving ease. Because some traits such as milk yield can only be measured in females, progeny testing is carried out to collect phenotypic data to calculate accurate EBVs for potential elite sires. Progeny testing involves breeding from a cohort of young potential sires that have been bred from the best performing dams, and then using the phenotypic records of the bull's daughters in the BLUP evaluation. This process takes approximately five to seven years from conception of potential sires to first proof.

Despite the improvement achieved using BLUP methods to predict EBVs, there are some disadvantages to the method (Calus, 2010). It is not possible to estimate accurate breeding values for an individual when no phenotypic information is available for either the animal itself or close relatives, and it could result in increased rates of inbreeding as the process inherently favours the selection of close relatives (Calus, 2010).

1.2 Use of molecular genetic information to improve evaluations

Steps were taken to try to address these issues, via mapping of quantitative trait loci (QTL) with subsequent application of marker assisted selection (MAS) (Dekkers and Hospital, 2002), where markers in close linkage disequilibrium (LD) with QTL were

used to improve predictions of merit before phenotypic information became available. A number of QTL associated with various production traits were successfully mapped, but MAS techniques were not as successful as hoped. This was because the mapped QTL did not explain a sufficient proportion of the total genetic variance for the traits, and also because the LD observed between the markers and QTL did not persist across families (Hayes and Goddard, 2003).

1.3 Genomic selection

It was suggested by Haley and Visscher (1998) that marker assisted selection techniques could be implemented at a genome-wide level, and in 2001, Meuwissen et al. (2001) described methodology to simultaneously estimate the effect of a dense panel of markers across the genome. This genome-wide selection technique would bypass the need for QTL mapping, and still allow an accurate prediction of the total genetic variance of a trait from the genome (Calus, 2010). This technique has since become widely known as genomic selection. The theory behind genomic selection is that every QTL affecting a trait is expected to be in LD with at least one SNP marker across a population, and so by simultaneously estimating SNP effects across the genome using a reference population containing individuals with both genotypic and phenotypic data, it is possible to estimate the genetic merit of individuals with no phenotypes, expressed as a genomic breeding value, or GEBV. When the Meuwissen paper was published back in 2001, the theory could only be tested using simulated data, as genotyping technologies were not yet advanced enough to produce dense marker maps. It wasn't until the release of high-throughput SNP genotyping chips

(Matukumalli et al., 2009) that the technique could be implemented in livestock populations.

The correlation between estimated breeding value and true breeding value is a measure of the accuracy of both traditional and genomic evaluations. The accuracy of EBVs of young bulls before progeny testing based on parental average is around 54%, after progeny testing is around 75% rising to around 90% accuracy for a proven bull (VanDoorMal, 2009). Selection decisions are generally made after progeny testing because the response to selection is dependent on the accuracy of EBVs, and using older animals with highly accurate EBVs gives a higher rate of genetic progress than using young bulls in the short term. The accuracy of GEBV for a young bull is approximately 75%. This level of accuracy makes the selection of young bulls more attractive, as the increase in the rate of response to selection achieved by making selection decisions at a younger age (and thereby shortening the generation interval) is greater than the loss due to selecting individuals with lower accuracy GEBVs (Lillehammer et al., 2011). The opportunity to use bulls at a younger age is also potentially cost saving as fewer bulls will need to be progeny tested (Schaeffer, 2006). Genomic selection has therefore become an extremely attractive prospect for the dairy breeding industry.

1.4 Genomic selection methodology

Genomic selection can be implemented using a wide range of different statistical methodologies. Models such as GBLUP (Meuwissen et al., 2001) have been adapted from traditional pedigree based BLUP evaluations in order to incorporate genomic data. GBLUP equates to the original BLUP model described by (Henderson, 1975),

but replaces the numerator relationship matrix **A**, with the genomic relationship matrix **G**, where relationships between individuals are calculated based on marker data rather than pedigree. Another commonly used method is SNP-BLUP, which can be considered equivalent to GBLUP, but fits a vector of individual SNP effects within the BLUP model as opposed to individual animal effects in GBLUP. Both models assume that SNPs are normally distributed with equal variance, and all SNPs have an effect on the trait of interest.

A large number of Bayesian methodologies have also been developed specifically for genomic selection, in order to make best use of any prior information that may affect the trait of interest. The main difference between the different Bayesian models is the prior distribution of SNP effects. In models such as BayesA (Meuwissen et al., 2001), all SNPs are assumed to have an effect on the trait of interest, whereas models such as BayesB and BayesC assume only a proportion of SNPs affect the trait of interest. A number of different parameters are incorporated into the priors, which can either be set to specific values, or estimated from the data itself (Nadaf et al., 2012). GBLUP models are more commonly used in commercial applications of genomic selection with large datasets, as they are less computationally intensive than Bayesian methods. However, the accuracies of GEBVs obtained using GBLUP are not sensitive to the number of QTL affecting the trait (Daetwyler et al., 2010), and so it has been suggested that Bayesian methods of evaluation may perform better for traits that are controlled by fewer QTL of larger effects (Hayes et al., 2009b).

The methods described above rely on a reference population that has both phenotypic and genotypic data. This data is often available for far fewer individuals than are

currently used for traditional genetic evaluations. The implementation of genomic selection therefore usually involves carrying out a traditional genetic evaluation, and then using de-regressed EBVs or daughter yield deviations (or yield deviations in the case of cows) as a "pseudo-phenotype" to compute genomic evaluations. A method has been suggested that allows us to bypass this extra step, and calculate GEBVs in a "single step" evaluation procedure, by combining information from the numerator relationship matrix **A**, and the genomic relationship matrix **G**, into an **H** matrix that can be incorporated into the mixed model equations described by Henderson (Misztal et al., 2009). This method allows the incorporation of records from animals without genotypes, and is known as single-step GBLUP, or HBLUP. Several studies have reported higher accuracies of evaluation for HBLUP than for other methods of prediction, which is likely a result of being able to considerably increase the size of the reference population.

1.5 Applications vs. theoretical expectations

The original paper by Meuwissen et al (2001) suggested that genomic selection methods could predict GEBVs with accuracies of up to 0.8. The concept has now been implemented in a number of species, such as cattle (Hayes et al., 2009b; VanRaden et al., 2009), pigs (Cleveland et al., 2013; Ibáñez-Escriche et al., 2014), poultry (Wolc et al., 2011; Wang et al., 2013) and sheep (Daetwyler et al., 2012b) across a range of different traits. The results of these studies have been variable, with the accuracy of evaluation ranging from approximately 0.35 to 0.95 (Cleveland et al., 2013). The dairy sector has seen the highest levels of uptake of genomic selection, which is mainly due to the structure of the industry, which relies on most

selection being carried out on sires after progeny testing. As a consequence of artificial insemination procedures, elite dairy bulls can have thousands of daughters that go into national milking herds. Information from all of these daughters is incorporated into genetic evaluations of these bulls, which results in highly accurate breeding values being estimated. These bulls can be used in a reference population for genomic evaluations, where the de-regressed EBV is equivalent to a phenotype with a heritability equal to the reliability of the EBV (Garrick et al., 2009). Studies using both simulated and empirical data have demonstrated that genomic selection accuracy is greater for traits with higher heritabilities (Calus et al., 2008; Luan et al., 2009; Daetwyler et al., 2008; de Roos et al., 2011), and so the use of high accuracy EBVs in genomic evaluation also leads to highly accurate GEBVs. Cattle also have a long reproductive cycle, and genomic selection can therefore also lead to large economic gains by reducing the generation interval (Jonas and de Koning, 2015; Lillehammer et al., 2011).

With regards to the performance of specific statistical methodologies, in studies where multiple models have been directly compared, no single method has been shown to consistently outperform others for all traits. In some cases, GBLUP has performed slightly better than Bayesian models (Daetwyler et al., 2012b; Luan et al., 2009), and in other cases the opposite is true (Hayes et al., 2009a). Where used, HBLUP has outperformed GBLUP and Bayesian models (Koivula et al., 2012; Gao et al., 2012; Mucha et al., 2015), and this may be due to the large increase in size of the reference population achieved by including animals without genotypes.

The results of genomic selection methods applied suggest that the following factors are central to success of the technique;

1. The observed level of LD between markers and QTL
2. The size of the reference population
3. The heritability of the trait of interest
4. The level of relationship between the reference population and the selection candidates.

The success of genomic selection relies on the persistence of LD between markers and QTL across the population of interest (Goddard and Hayes, 2007). The proportion of QTL variance explained by markers can be expressed as an r^2 value. It has been suggested that the level of LD should be $r^2 = \geq 0.2$ to be able to compute GEBVs with an accuracy of around 0.8 (Calus et al., 2008; Goddard and Hayes, 2007; Meuwissen et al., 2001). The average r^2 value between markers and QTL should increase as marker density increases, and so higher density genotypes should yield more accurate GEBVs (De Roos et al., 2008). LD is related to effective population size (Ne), in that the average level of LD between loci decreases as Ne increases. The accuracy of genomic selection therefore has an inverse relationship with population Ne, and so populations with higher Ne (e.g. sheep) will demonstrate lower genomic selection accuracies than populations with a low Ne (e.g. dairy) at a given reference population size.

The number of animals included in the reference population appears to be one of the main limiting factors of genomic selection studies based on empirical data (Hayes et al., 2009b; VanRaden et al., 2009). The number of observations for each SNP allele will increase as the size of the reference population increases (Hayes et al., 2009b). As mentioned previously, the heritability of the trait of interest is also important, as better accuracies are generally observed for traits with a higher heritability (Calus et al., 2008; Luan et al., 2009) and it has been shown by Daetwyler that accuracy is directly proportional to the heritability of the traits of interest in the reference population (Daetwyler et al., 2008). Hayes et al. (2009a) have also observed an inverse relationship between heritability and the size of the reference population.

The relationship between the reference population and the selection candidates also has an effect on accuracy, as animals that are closely related to those in the reference population have been observed to have higher accuracy GEBVs (Clark et al., 2012; Misztal, 2011).

1.6 Extending genomic selection across breeds

Countries including the UK and the USA are currently implementing genomic selection within the Holstein breed, and also other dairy cattle breeds such as the Jersey and Guernsey. These evaluations are carried out using data from the Illumina BovineSNP50 bead chip. There is a great deal of interest across the livestock breeding sectors in “across breed evaluations”, a term which encompasses a number of ideas, including the potential for the use of multi-breed reference populations to translate information gained from numerically large breeds to improve predictions in

smaller breeds, and selection of animals for crossbred performance (see Figure 1.1). The use of a crossbred reference population to predict into a crossbred validation population has been successfully implemented in goats (Mucha et al., 2015).

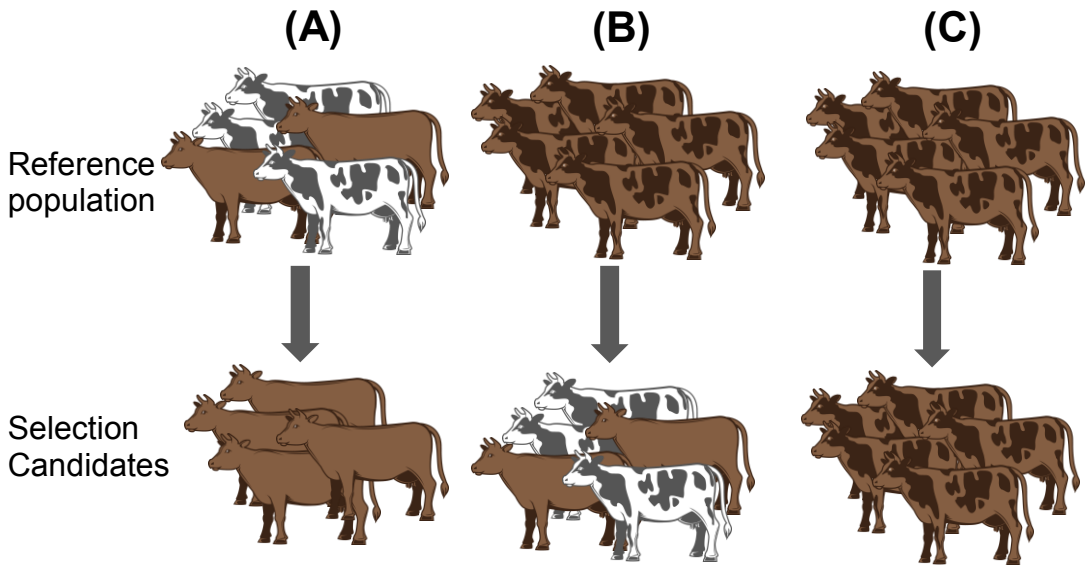

Figure 1.1 Shows A) a multi-breed evaluation model, where a reference population comprised of two or more pure breeds is used to calculate genomic evaluations in one of the breeds represented in the reference population, B) a crossbreed evaluation model, where a reference population of crossbred animals is used to calculate genomic evaluations for purebred animals to be used for crossbred performance, and C) a crossbreed evaluation model where a reference population of crossbred animals is used to calculate genomic evaluations for crossbred selection candidates.

1.6.1 Multi-breed evaluations

The previously mentioned factors affecting accuracy are relevant for both within breed and multi-breed predictions. However, for multi-breed prediction to be successful, it is also necessary for the LD phase to persist between markers and QTL across the population of interest, i.e. the same marker should be inherited with the QTL across all breeds in the reference population (Goddard and Hayes, 2007).

A number of studies have been carried out using the Illumina 50k chip to assess the potential benefits of using multi-breed reference populations in genomic evaluations (Hayes et al., 2009a; Karoui et al., 2012; Kizilkaya et al., 2010; Olson et al., 2012; Pryce et al., 2011) Although the majority of these studies did report an increase in prediction accuracy when using a multi-breed reference population compared to a purebred reference population, these gains were relatively small and were inconsistent across breeds and traits. The consensus from these studies was that a higher density set of markers would be required to successfully compute across breed genomic evaluations (Kizilkaya et al., 2010; De Roos et al., 2009). De Roos et al. (2008) investigated LD and persistence of phase in Holstein Friesian, Jersey, and Angus cattle, and concluded that around 300,000 SNP markers would be needed to find markers in LD with QTL across breeds (De Roos et al., 2008).

Wientjes et al (2013) investigated the effect of factors such as LD and familial relationships on the reliability of genomic evaluations, and discovered that reliability of GEBVs is largely due to the presence of familial relationships between individuals in the reference population and selection candidates, rather than the persistence of LD. Other studies have also highlighted the importance of relationships between the reference population and the selection candidates when looking to improve accuracy (Daetwyler et al., 2012a; Clark et al., 2012). It is likely that the lack of familial relationships across pure breeds is one of the limiting factors in implementing successful multi-breed evaluations.

In 2010, Illumina released the BovineHD SNP chip, which contains approximately 777,000 SNP markers across the genome, decreasing the average inter-marker distance from around 55kb to around 4kb. It was hoped that the markers on the HD chip would be dense enough to pick up LD prior to breed divergence, and improve the accuracy of across breed evaluations. However, studies using the HD chip have reported limited improvements in accuracy (Erbe et al., 2012; Ertl et al., 2014; Harris et al., 2011) and the uptake of the HD technology has been low. Of importance is that the increase in accuracy in these studies has been lower in Holstein populations than in other populations, for example the Jersey, and Red Dairy Cattle (Erbe et al., 2012; Su et al., 2012). This suggests that the increase in marker density may be of more use in populations with a higher Ne, as the relative increase in average LD between markers would be higher than in populations with low Ne (Su et al., 2012). Though the use of the HD chip does increase the average pair-wise LD between markers, it is possible that the increase in accuracy is lower than expected due to the larger number of parameters to be estimated, and also that the models used currently to calculate genomic predictions are suboptimal (Su et al., 2012) .

As previously mentioned, a number of different density SNP chips have been tested for their use in genomic selection techniques. To allow widespread commercial application of genomic selection, a balance needs to be struck where the number of SNPs used maximises the accuracy of evaluation, but the cost per genotype does not prohibit the creation of a large reference population. The low increases in accuracy when using HD genotypes (Erbe et al., 2012; Ertl et al., 2014; Harris et al., 2011) along with the higher cost per genotype, have resulted in a reluctance to adopt

genotyping using the HD chip. An alternative to directly genotyping at HD would be to impute up from lower density panels using a reference panel of animals with HD genotypes. A number of software packages are now available for genotype imputation, and genotypes can be imputed with accuracies of up to 99% dependent on the difference in density between the chips being imputed and the imputation software used (Berry et al., 2014; Hozé et al., 2013; Pimentel et al., 2015).

As well as chips for hundreds of thousands of markers now being available for genotyping, the cost of carrying out whole-genome sequencing on samples has dropped significantly in recent years, leading some to suggest that sequence data could be a feasible source of data for use in genomic evaluation, as there would no longer be any need to rely solely on LD between markers and QTL, as causal variants should presumably be present (Meuwissen and Goddard, 2010). Data can either be collected via next-generation sequencing technologies, or imputed from genotype data. Iheshiulor et al. (2016) used a simulated data set to show that using whole-genome sequence data results in a higher evaluation accuracy than SNP panels when using a multi-breed reference population, but it is yet to be seen whether this advantage will translate into empirical data.

The UK dairy industry is extremely interested in the potential for carrying out multi-breed predictions. The UK launched commercial genomic evaluations for Holstein cattle in 2012, but has so far not been able to offer them for other breeds (such as the British Friesian and the Guernsey), as the breeds are too numerically small to construct a single-breed reference population of sufficient size to accurately predict

GEBVs. Breeding companies and breed societies in the UK are keen to see genomic evaluations available for further breeds, and so incorporating genomic information from multiple breeds into a single reference population may be a way to facilitate this.

1.6.2 Crossbred evaluations

Genomic selection is also of interest to allow accurate selection of crossbred animals, or of purebred animals for crossbred performance. If this is to be achieved by incorporating crossbred genotypes into the reference population, further issues are faced due to the effects of heterosis through dominance and epistasis, which are currently ignored when calculating GEBVs within breeds. These non-additive genetic effects are likely to have a larger effect on the phenotype in crossbred animals than purebred animals, and using purely additive genetic models to calculate predictions may result in bias (Toosi et al., 2010).

Work has been carried out to determine the importance of breed of origin when estimating GEBVs for crossbred animals. Simulation studies suggest that crossbred animals can be used in predictions for crossbred performance without taking into account breed of origin (Ibánez-Escriche et al., 2009; Toosi et al., 2010). Ibañez-Escriche et al suggested that taking into account breed specific SNP effects may improve accuracy when the breeds in question are distantly related, or if sufficient records are available to accurately estimate breed-specific SNP effects (Ibánez-Escriche et al., 2009). A study by Makgahlela et al. used both GBLUP and a multi-trait random regression model that takes account of breed specific marker effects to

predict GEBVs for milk traits in Red Dairy Cattle (Makgahlela et al., 2013). The model accounting for breed specific allele effects resulted in a small gain in accuracy for some traits, but accuracies achieved were nevertheless low. A simulation study by Zeng et al. has demonstrated the use of a model including dominance effects when selecting purebreds for crossbred performance, which was shown to perform better than a breed specific SNP model (Zeng et al., 2013). Studies based on empirical data have mainly been carried out on pigs, as crossbreeding is an essential part of the pig breeding industry. Xiang et al (2016) demonstrated that GEBVs for crossbred performance could be estimated with reliabilities of between 0.26 and 0.39 for purebred boars using the "single step" method put forward by Christensen (Christensen et al., 2014). Of note was that GEBV reliability was improved when crossbred genomic information was incorporated into the model.

Work has also been on-going to test the potential for using crossbred reference populations to predict the merit of crossbred offspring, particularly in beef cattle and pigs where there is a focus on carcass trait performance in the crossbred slaughter generation, as well as in UK crossbred dairy goats (Mujibi et al., 2011; Vallée et al., 2014; Mucha et al., 2015). As is the case in within-breed genomic selection, the size of the reference population is still paramount, but in populations where traditional BLUP EBVs were also available for the selection candidates (Mujibi et al., 2011; Mucha et al., 2015), the accuracy of GEBVs calculated was not high enough to outweigh the costs associated with genotyping.

Approximately 50% of the beef produced in the UK is not from traditional beef breeds, but in fact comes as a by-product of the dairy industry, mainly in the form of bull calves born to cows in the milking herd that are of no use for breeding. It has become common practice to mate dairy cows with beef bulls to improve the quality of the resulting calves, so that they are more suitable for finishing. The quality of calves is assessed at birth, with those designated as unsuitable for finishing being culled. The availability of genomic evaluations for crossbred cattle in the UK would allow for the genetic improvement of dairy-beef crosses, which in turn would have a positive impact on both the efficiency of the farms, and also animal welfare by reducing the rate of culling at birth.

1.7 Thesis outline and main objectives.

Genomic selection has become a widely used commercial tool for within-breed evaluations in the dairy industry; however, the application of the technique in "across-breed" scenarios has not yet been successful. This is likely to be due to a combination of factors such as relatively low numbers of phenotypic records, insufficient marker density, low numbers of crossbred animals genotyped and the lower levels of relationship between reference and validation individuals in an across breed population.

The aim of this PhD is to investigate the potential for successfully implementing genomic selection in across-breed scenarios and to understand the prerequisites for constructing a genomic selection programme that utilises the most genomic information from all sources.

Chapter 2 concerns the multi-breed genomic selection scenario, where data from a larger breed (Holstein) is incorporated into a multi-breed reference population to allow evaluations of a numerically smaller breed, the British Friesian.

Chapter 3 concerns the crossbred genomic selection scenario in the context of a crossbred dairy cattle population from a developing country, using high-density genotype data along with comparing multiple genomic selection models.

Chapter 4 considers the use of whole genome sequence data for genomic selection, and investigates the possibility of extracting pertinent SNP information from whole genome sequence in order to estimate genomic relationships between individuals.

Chapter 2: Using multi-breed genomic evaluation to improve genomic prediction in a numerically small population.

2.1 Introduction

Genomic selection has been widely adopted in the dairy cattle industry, with commercial genomic evaluations now being published for a number of traits and breeds worldwide. However, the success of genomic selection depends on being able to predict breeding values with a high enough level of accuracy, and to achieve this a large reference population of animals with both phenotypes and genotypes is needed (Hayes et al., 2009b; VanRaden et al., 2009).

There are a number of dairy breeds where the genotyped population size is too small to create a single breed reference population of sufficient size to allow the prediction of accurate genomic breeding values (Thomasen et al., 2014). The British Friesian is one such breed, since there are only a few thousand milking Friesian cows in the UK. The full complement of Friesian genotypes available in the UK and Ireland totals just 98, and it is unlikely that this number will increase dramatically due to small overall population size.

The Holstein and Friesian breeds both originated in the Netherlands and were exported to UK and USA in the late 1800's. The two breeds are closely enough related that they are often collectively known as Holstein-Friesian. Selection in the USA has been for high milk yield whereas in the UK both meat and milk were selected for simultaneously (Mingay, 1982), leading to populations of dairy cows with distinctly different phenotypes. However, since the early 1980's selection in the UK has mirrored that of the USA and semen used in the UK has been imported from the USA or other countries where US bulls have been extensively used. Thus the

populations have converged from a genetic viewpoint. However, for the purposes of genetic evaluation in the UK they are evaluated together, but the breeds are considered separate and results are published on breed specific bases.

There are two scenarios in which genomic evaluations for Friesians could be attempted with the current number of genotypes available. The first of these would be to ignore the breed status of Friesian animals, and use the current Holstein training population to predict their GEBVs. The second would be to include the limited number of Friesian bull genotypes available into a multi-breed training population along with the Holstein genotypes. It is hypothesised that higher accuracies of evaluation for Friesians would be obtained if Friesian genotypes were to be incorporated into a multi-breed training population. Indeed, previous studies have shown higher accuracy from a multi-breed training population, compared with using one pure breed to predict another (Hayes et al., 2009a; Brøndum et al., 2011; Zhou et al., 2014a).

Another question to pose is which model is most suitable for the analysis of multi-breed genomic data. The current Holstein genomic evaluations in the UK use a standard SNP-BLUP model to estimate GEBVs. A number of different methods have been suggested for a multi-breed analysis (Erbe et al., 2012; Weber et al., 2012; Carillier et al., 2014; Hozé et al., 2014) with Bayesian methods of estimation giving slightly better estimates than GBLUP (Hayes et al., 2009a; Zhou et al., 2014a), but this is not completely consistent across populations and traits. In addition, there is a wealth of recorded data available for Friesian bulls without genotypes, and so

another method of interest would be the single step method (Legarra et al., 2014), as a technique which allows us to make full use of the data available and may result in improved prediction accuracy.

As well as the size of the training population, another important factor underlying the success of genomic selection is the persistence of LD between marker and QTL in general and much more so in an across-breed situation (Goddard and Hayes, 2007). The Illumina 50K Bovine chip has an average interval between markers of around 67kb (Matukumalli et al., 2009), and previous studies have suggested that a higher density of markers may be necessary to successfully compute evaluations across breeds (Kizilkaya et al., 2010; De Roos et al., 2008). Imputation of genotypes to high density (777k) before analysis may be a solution to this problem, but the extra markers may also be a source of increased noise in the data considering the limited number of Friesian genotypes available.

The present study therefore has three distinct objectives;

a) To compare the accuracy of a multi-breed vs Holstein only reference population in Friesian predictions
b) To assess the accuracy of two different methods of GEBV estimation in a multi-breed setting
c) To assess the utility of using HD genotypes vs 50k genotypes for multi-breed genomic evaluations

2.2 Methods

2.2.1 Data

The initial dataset comprised 21,646 Holstein and 23,003 Friesian bulls born between 1960 and 2008 that had de-regressed estimated breeding values (dEBVs) for three production traits (milk, fat and protein yield) and two non-production traits (lifespan and milk somatic cell count). Of these bulls, 21,646 purebred Holsteins and 98 purebred Friesians also had whole-genome genotype data available. A subset of approximately 4,000 Holsteins were selected for inclusion in the analysis, to a) ensure that the Friesians made up over 1% of the full population, and b) to ensure that enough individuals were present in the reference population to predict GEBVs at a reasonable level of accuracy. As the pedigree relationship between individuals in the reference population has an impact on evaluation accuracy (Wientjes et al., 2013), the average relationship of each Holstein animal with the Friesian animals was calculated based on the numerator relationship matrix, and those animals with the highest average relationship to the Friesian animals in analysis were taken forward for analysis.

In all cases, animals with conflicts between pedigree and genotype were discarded. A cut off birth year of 1996 was used to divide the population into reference and validation sets. This cut off year is much earlier than would be used in a commercial evaluation, but the Holstein animals were generally younger than the Friesian animals, and this was the most appropriate point to divide the dataset in a way that would allow a reasonable number of Friesians in both the training and validation populations.

Preliminary analyses revealed that genotyping errors had arisen for 15 Friesian animals as part of the recoding process. These genotypes were removed from further analyses. Along with uncovering genotyping errors, preliminary analyses also showed that 29 of the remaining 83 Friesians were in fact Holstein-Friesian crosses as opposed to purebred Friesians. These 29 crosses were still included in the analysis but not included in either of the purebred validation populations.

The final data set therefore consisted of 4,468 genotyped individuals; 4,385 Holsteins, 54 Friesians and 29 Holstein-Friesian crosses. All Friesian and Holstein-Friesian cross animals in the dataset were genotyped using the Illumina BovineHD (777k) chip, along with 330 Holsteins, and so imputation to HD was carried out for those Holstein animals genotyped with either version of the Illumina 50k chip. Imputation was carried out using findhap (VanRaden et al., 2013), with 1,725 Holstein HD genotypes used for population haplotyping. Of the 4,055 Holsteins that needed to be imputed to HD, 19 could not be successfully imputed as no pedigree data was available, and so these genotypes were removed from the analysis, leaving a set of 4,363 genotyped Holsteins, 54 genotyped Friesians, and 29 genotyped Holstein-Friesian crosses (total 4,449).

Raw genotype data along with chromosome and map position was provided for analysis. SNPs were edited by loci, and SNPs with a call rate <0.95 or a minor allele frequency <0.05 were removed, along with SNPs on the X chromosome and those out of Hardy-Weinberg equilibrium. This left us a panel of 536,229 markers for analysis, 36,666 of which were common to the 50k chips.

Table 2.1 Numbers of Holstein (Hol), Friesian (Fri) and Holstein-Friesian cross (HFX) animals used in analysis, separated into reference and validation populations (animals in the reference population were born in or before 1996, animals in the validation population were born in or after 1997).

		Reference Population			Validation Population		
Trait	Data Type	Hol	Fri	HFX	Hol	Fri	HFX
Milk yield	Genotyped	2,391	36	24	1,971	18	5
	Ungenotyped	2,496	19,651	1,334	0	0	0
Fat yield	Genotyped	2,392	36	24	1,971	18	5
	Ungenotyped	2,500	21,846	1,475	0	0	0
Protein yield	Genotyped	2,392	36	24	1,971	18	5
	Ungenotyped	2,500	21,821	1,474	0	0	0
Lifespan	Genotyped	2,384	22	22	1,892	17	4
	Ungenotyped	2,362	4,806	431	0	0	0
Somatic Cell	Genotyped	2,353	20	20	1,968	18	4
Count	Ungenotyped	2,590	1,361	213	0	0	0

2.2.2 Formation of G matrix and Principal component analysis (PCA)

Relationship matrices based on pedigree data (**A**) and SNP genotype data (**G**) were calculated for the full set of genotyped animals for both analyses. The **G** matrix was calculated using VanRaden's first method(VanRaden, 2008), where $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$, where $\boldsymbol{Z}$ is a design matrix of centred genotypes, and p_i is the allele frequency estimated across breeds for the major allele at SNP i. Both the **A** and the **G** matrices were calculated using the program preGSf90 (Misztal et al., 2002). Principal component analysis was then performed on each of these matrices using the R function "princomp" (R Core Team, 2013), and the resulting principal components plotted to determine sources of variation between individuals.

2.2.3 Comparison of G matrices computed from different SNP densities

The availability of HD genotypes allowed us to calculate two **G** matrices for the population, the first of these based on all SNPs that passed quality control ($\mathbf{G_{HD}}$), and the second based only on SNPs common to the 50k chip ($\mathbf{G_{50k}}$). The two **G** matrices were compared by computing the correlation between all elements of **G**, and calculating the regression coefficient.

2.2.4 Linkage disequilibrium

LD between markers and QTL is vital for the success of genomic selection techniques (Goddard and Hayes, 2007). The level of LD between two loci decreases as the distance between the loci increases, as the further apart the two loci are, the higher the probability of a recombination event happening between the two loci during meiosis. As breeds diverge over many generations, further recombination events will take place, degrading the level of LD between more distant loci in the process. We were therefore interested in measuring the persistence of LD across the genome within this population of cattle. PLINK software (Purcell et al., 2007) was used to estimate the level of LD between SNP pairs up to 1Mb apart on the same chromosome for all genotyped animals. LD was calculated as the squared correlation (r^2) of alleles at two SNP loci across the population. Within breed LD was also calculated for Holstein and Friesian animals. LD calculations were carried out on both the higher and lower density sets of markers after sample QC.

2.2.5 Study design

GEBVs were calculated using three reference populations, (i) Holstein only, (ii) Friesian only and (iii) a combined reference population containing all Holsteins, Friesians and Holstein-Friesian crosses born before 1997. Each of these was used to predict accuracies for Holstein and Friesian subsets of the full validation population.

2.2.6 Statistical methods

2.2.6.1 GBLUP

A univariate GBLUP model was used to estimate GEBVs using de-regressed estimated breeding values (dEBVs) as phenotypes for each of the five traits. The de-regression was carried out on the official UK proofs for December 2014 using national parameters. The BLUPF90 software package (Misztal et al., 2002) was used to fit the following mixed model: $\mathbf{y = Xb + Za + e,}$ where **y** is a vector of dEBVs, **b** is a vector of fixed effects consisting of the mean and breed, **a** is a vector of animal effects, and **e** is a vector of residual effects; **X** and **Z** are respective incidence matrices. The distribution of **a** was assumed to be multivariate $N(0,\mathbf{G}\sigma^2_a)$, and the distribution of **e** was assumed multivariate normal $N(0,\mathbf{I}\sigma^2_e)$. All dEBVs were weighted in the model by the number of effective daughter contributions (EDC).

2.2.6.2 HBLUP

As indicated previously, although the number of Friesian genotypes available was small, EBVs were available for far more animals. The "single-step", or "HBLUP" method (Legarra et al., 2014), which enables data from animals without genotypes to

be incorporated into the reference population, was also tested. Data relating to dEBVs for ungenotyped animals born prior to 1997 was also available for up to 2,590 Holsteins, 21,846 Friesians and 1,475 Holstein Friesian cross bulls dependant upon the trait analysed (Table 2.2). This data was incorporated into an HBLUP analysis using an equivalent model to that shown above, but where the distribution of **a** was assumed to be multivariate $N(0,\mathbf{H}\sigma^2_a)$, where the inverse of **H** is calculated as

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G^{-1}} - \boldsymbol{A_{22}}^{-1} \end{pmatrix}$$

(Aguilar et al., 2010), where $\mathbf{A_{22}}$ relates to animals with genotype information. As in the GBLUP analysis, all dEBVs were weighted in the model by the number of effective daughter contributions (EDC).

2.2.7 Accuracy of genomic evaluations

For all three statistical methods, the accuracy of evaluation was calculated as the Pearson correlation coefficient between dEBV and GEBV. Accuracies were calculated separately for the Holstein validation population and the Friesian validation population. An estimate of the level of prediction bias was obtained using the slope of the regression of dEBV on GEBV. Differences in accuracies were tested for significance using the Fisher r-to-z transformation.

2.3 Results

2.3.1 Principal components analysis

Plots of principal components 1 and 2 based on the **A** matrix, the $\mathbf{G_{50k}}$ matrix and the $\mathbf{G_{HD}}$ matrix are shown in Figures 2.1 to 2.3. For all three matrices, principal component 1 related to variation between sire families, and accounted for 26.4% (**A**), 20.3% ($\mathbf{G_{50k}}$) and 18.7% ($\mathbf{G_{HD}}$) respectively. The second principal component related to variation between breeds, and accounted for 7.9% (**A**), 7.9% ($\mathbf{G_{50k}}$) and 15.7% ($\mathbf{G_{HD}}$) of the total variance respectively. While the **A** and, especially, $\mathbf{G_{HD}}$ plots show distinct clustering between Holstein and Friesian animals, all animals cluster together based on the $\mathbf{G_{50k}}$ matrix. Three Holstein animals clustered with the Friesians and Holstein-Friesian crosses based on the $\mathbf{G_{HD}}$ matrix. These animals are likely to be Friesian animals mislabelled as Holsteins, however, to eliminate any uncertainties regarding breed, the decision was made to remove these animals from the analysis. This left a total of 4,446 animals to be used in the estimation of GEBVs.

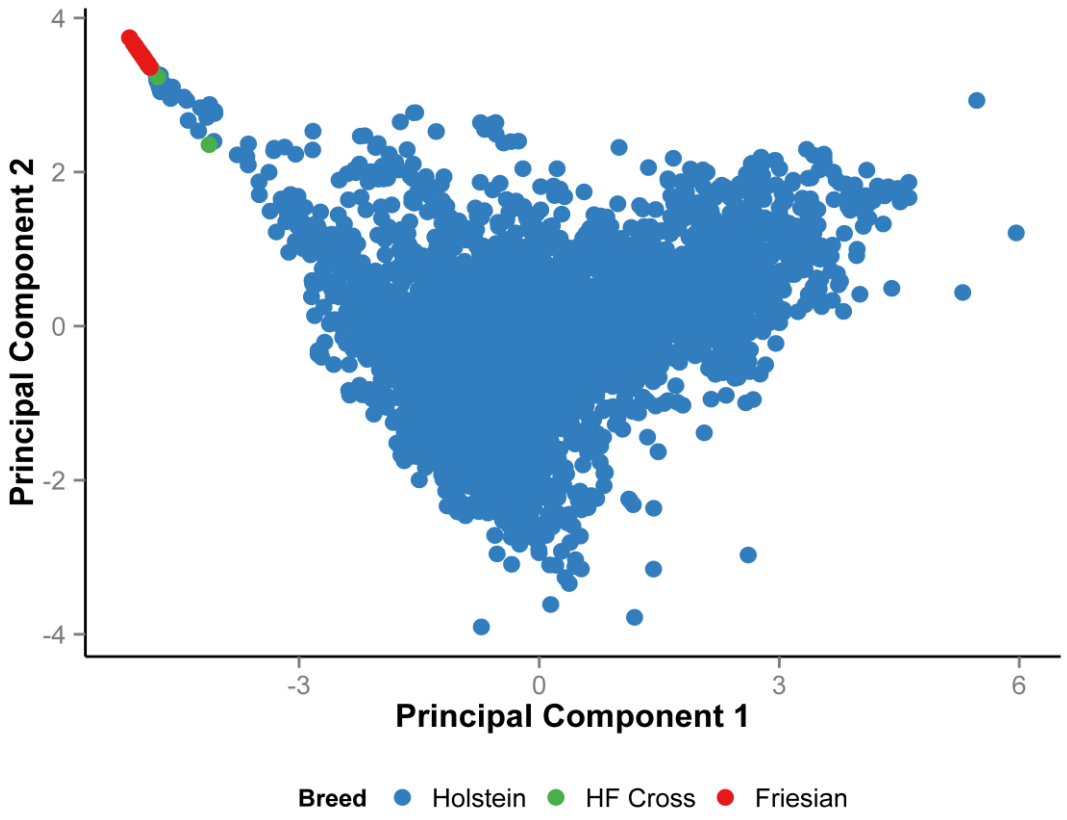

Figure 2.1 Principal components 1 and 2 based on a principal components analysis of the numerator relationship matrix **A**, with individuals coloured by breed.

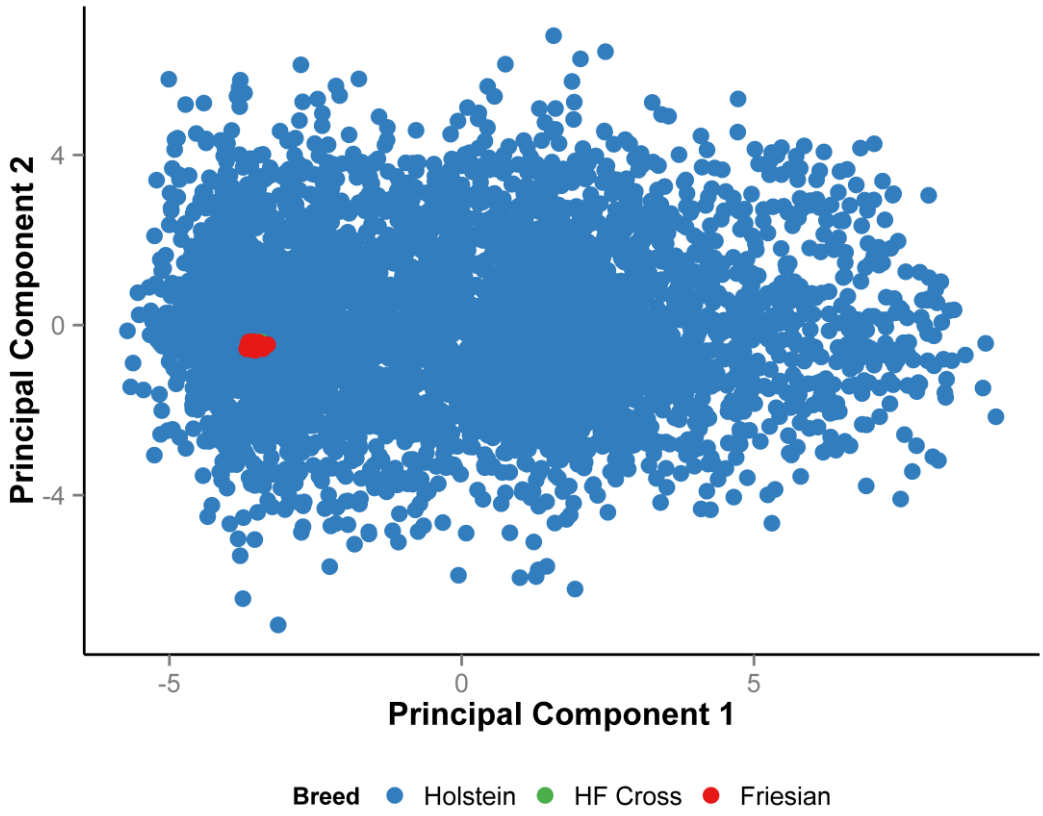

Figure 2.2 Principal components 1 and 2 based on a principal components analysis of the genomic relationship matrix $\mathbf{G}_{50k}$, with individuals coloured by breed.

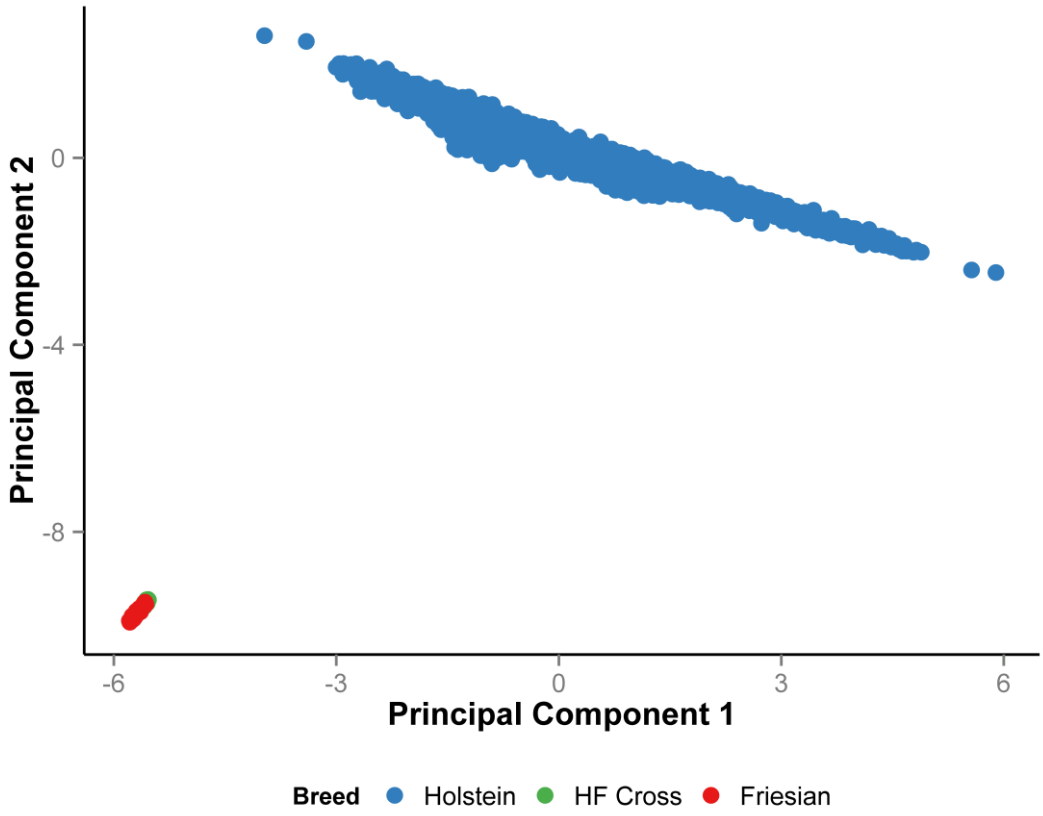

Figure 2.3 Principal components 1 and 2 based on a principal components analysis of the genomic relationship matrix $\mathbf{G}_{HD}$, with individuals coloured by breed.

2.3.2 Comparison between G Matrices

Figure 2.4 shows all elements of $\mathbf{G_{HD}}$ plotted against all elements of $\mathbf{G_{50k}}$. Correlation and regression coefficients were calculated for each breed relationship group (for example Holstein/Holstein relates to the relationship coefficient between two Holstein animals), and can be seen in Table 2.2.

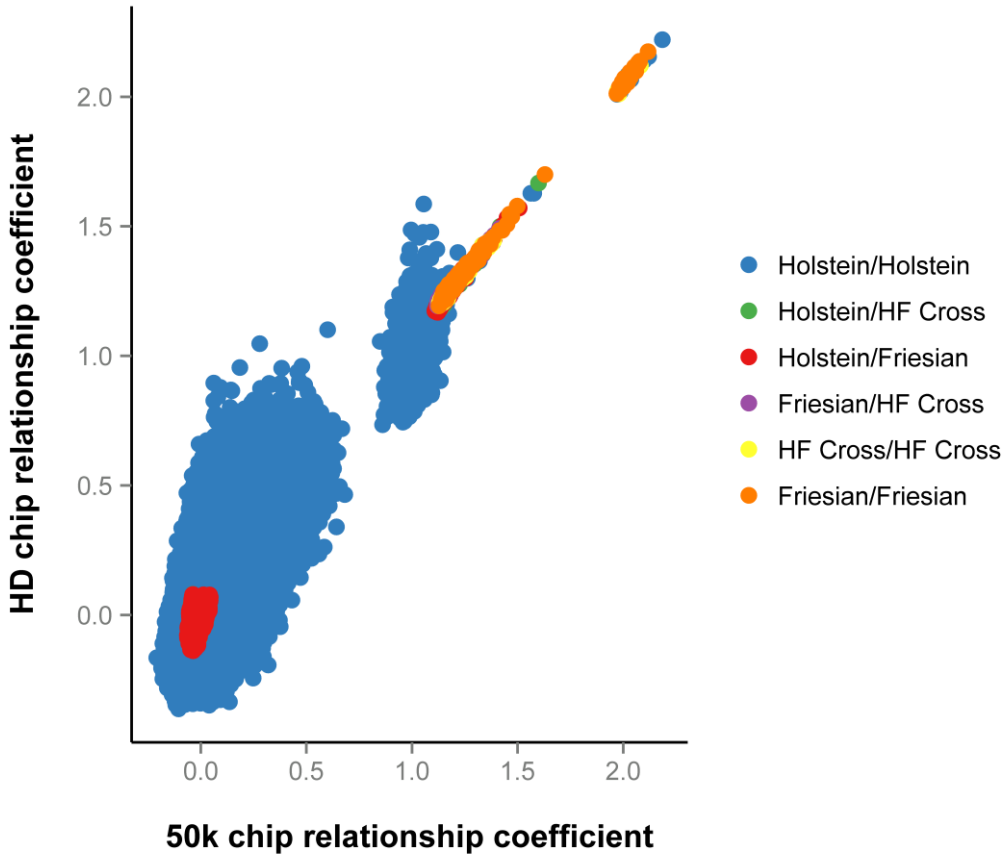

Figure 2.4 Relationship coefficients from the genomic relationship matrix $\mathbf{G_{50k}}$ plotted against relationship coefficients from the genomic relationship matrix $\mathbf{G_{HD}}$.

Table 2.2 Correlation (r) and regression (b) coefficients for the six breed relationship groups, between relationship coefficients from $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$.

Relationship	r	b
Holstein/Holstein	0.52	0.28
Friesian/Friesian	1.00	1.03
HF Cross/HF Cross	1.00	1.03
Holstein/Friesian	0.93	0.81
Holstein/HF Cross	0.92	0.81
Friesian/HF Cross	0.97	0.96

The correlation between relationship coefficients was above 0.90 for five of the six breed relationship groups, with the exception of Holstein/Holstein relationships. This was unexpected, and potentially due to the fact that a large proportion of the HD Holstein genotypes were obtained via imputation. To investigate this, **G** matrices were re-calculated for only those animals for which full HD genotypes were non-imputed. Correlation and regression coefficients were again calculated for each breed relationship group, and are shown in Table 2.3, with the associated correlation plot shown in Figure 2.5.

The correlation for Holstein/Holstein relationships between $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$ increased significantly (0.52 to 0.71, p = 0) when no imputed data was incorporated into **G** matrix calculations, with the regression coefficient also increasing from 0.28 to 0.87. The correlations between $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$ for all across-breed relationships were also significantly higher when no imputed data was used (p = 0).

The accuracy of imputation for 50k genotypes was estimated within findhap, and ranged from 97% to 99%.

Table 2.3 Correlation (r) and regression (b) coefficients for the breed relationship groups, between relationship coefficients from $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$.

Relationship	r	b
Holstein/Holstein	0.71	0.51
Friesian/Friesian	1.00	1.03
HF Cross/HF Cross	1.00	1.03
Holstein/Friesian	0.97	0.88
Holstein/HF Cross	0.97	0.88
Friesian/HF Cross	0.99	0.99

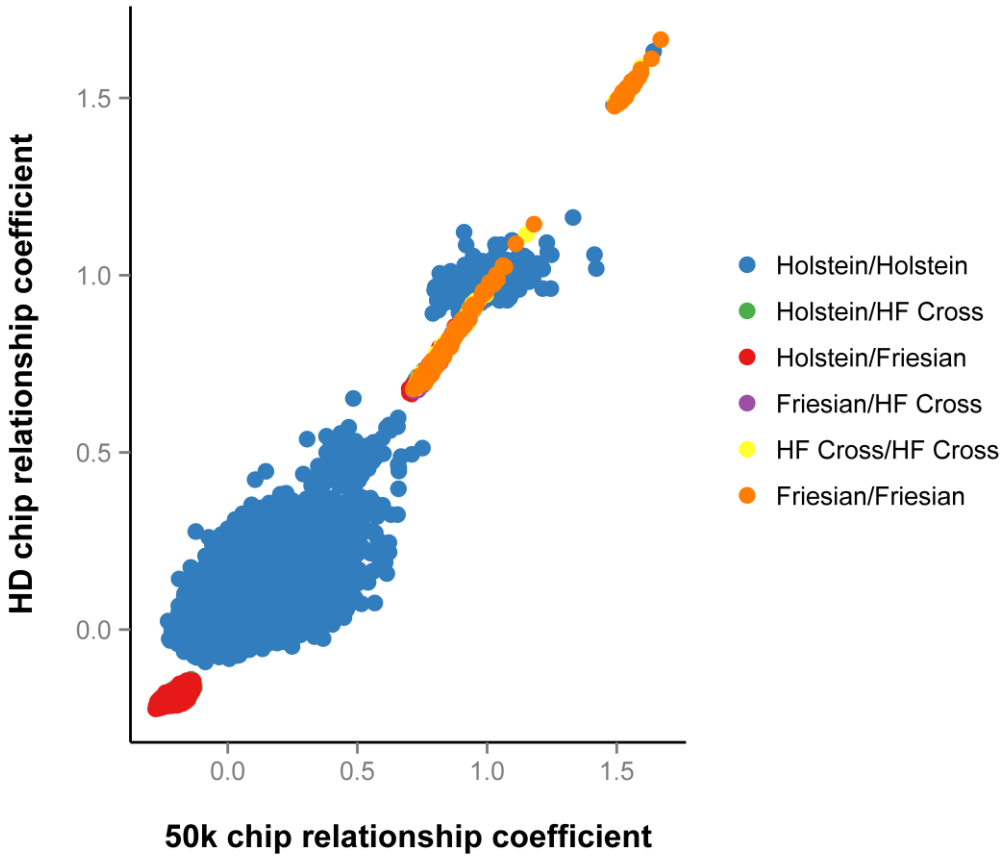

Figure 2.5 Relationship coefficients from the genomic relationship matrix $\mathbf{G}_{HD}$ plotted against relationship coefficients from the genomic relationship matrix $\mathbf{G}_{50k}$, based on non-imputed genotypes only.

At this point we also became interested in how closely the relationships calculated from genotypes ($\mathbf{G}_{50k}$ and $\mathbf{G}_{HD}$) matched relationships calculated from pedigree data (**A**). Each **G** matrix based on non-imputed data was therefore correlated with the numerator relationship matrix **A** for non-imputed animals, with regression coefficients also calculated. As in previous analyses, correlation and regression coefficients were calculated for all elements, as well as for each breed relationship group. All correlation and regression coefficients are shown in Table 2.4. The associated correlation plots are shown in Figures 2.6 and 2.7 respectively.

The correlations observed between **A** and $\mathbf{G_{50k}}$ and also **A** and $\mathbf{G_{HD}}$ were lower than the correlations observed between $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$. The overall correlation between **A** and $\mathbf{G_{HD}}$ was significantly higher than the correlation between **A** and $\mathbf{G_{50k}}$ for Holstein/Holstein relationships (p = 0).

Table 2.4 Correlation (r) and regression (b) coefficients for the - breed relationship groups, for relationship coefficients from; 1) **A** vs $\mathbf{G_{50k}}$, and 2) **A** vs $\mathbf{G_{HD}}$

	A vs $\mathbf{G_{50k}}$		**A** vs $\mathbf{G_{HD}}$	
Relationship	r	b	r	b
Holstein/Holstein	0.61	0.88	0.87	0.90
Friesian/Friesian	0.88	0.63	0.88	0.66
HF Cross/HF Cross	0.99	0.75	0.99	0.77
Holstein/Friesian	0.02	2.38	0.02	1.63
Holstein/HF Cross	0.01	0.18	0.00	0.05
Friesian/HF Cross	0.35	0.28	0.34	0.28

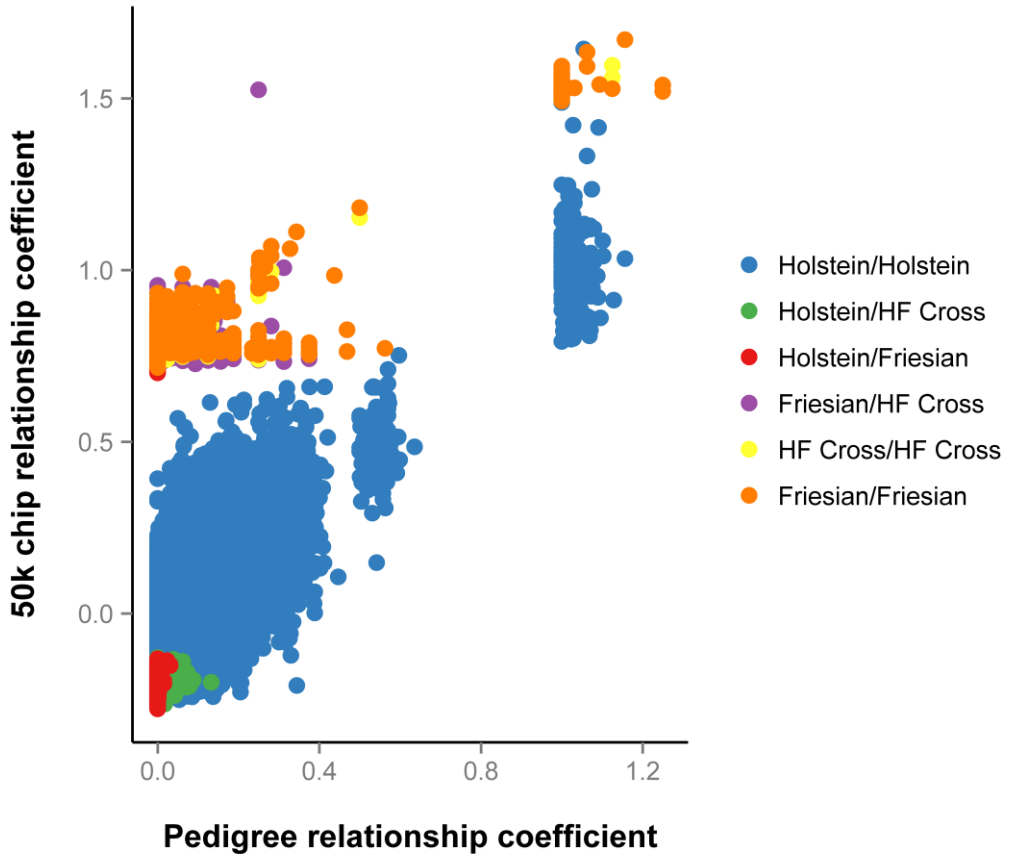

Figure 2.6 Relationship coefficients from the numerator relationship matrix **A** plotted against relationship coefficients from the genomic relationship matrix $\mathbf{G_{50k}}$, based on non-imputed genotypes only.

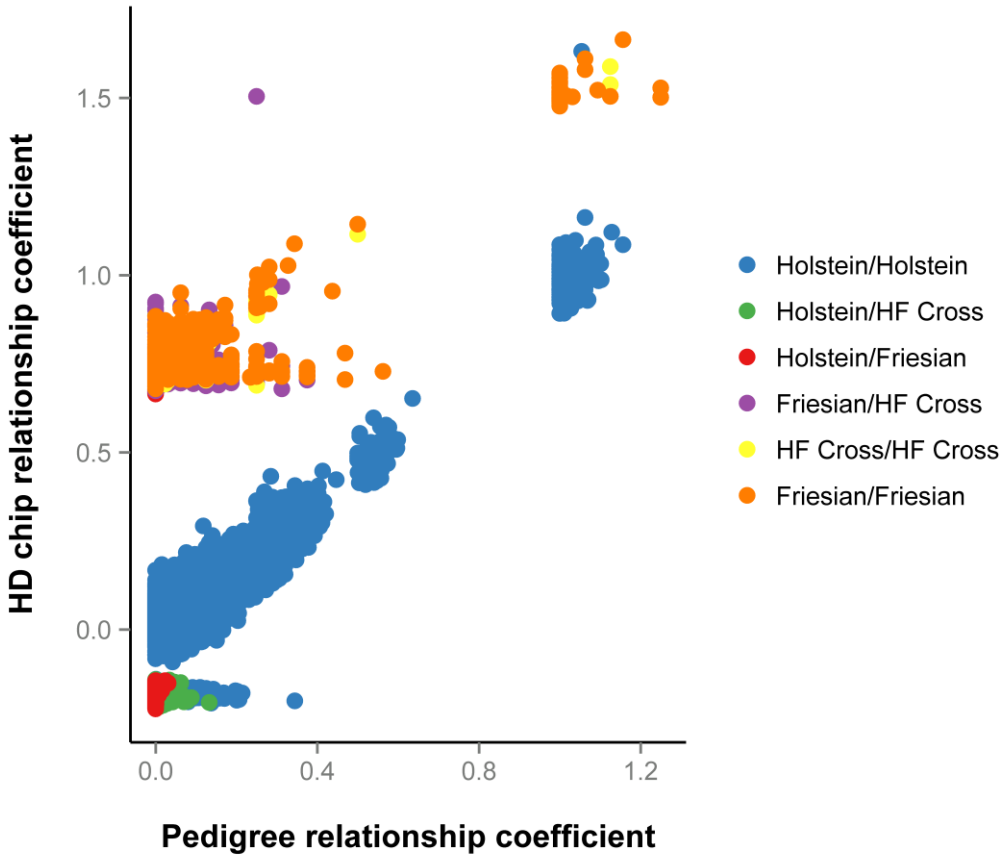

Figure 2.7 Relationship coefficients from the numerator relationship matrix **A** plotted against relationship coefficients from the genomic relationship matrix $\mathbf{G}_{HD}$, based on non-imputed genotypes only.

Along with repeating the correlation of **G** matrices, PCA analysis was also re-run on each of the $\mathbf{G}_{50k}$ and $\mathbf{G}_{HD}$ matrices calculated using non-imputed animals. Plots of principal components 1 and 2 for $\mathbf{G}_{50k}$ and $\mathbf{G}_{HD}$ can be seen in Figures 2.8 and 2.9 respectively. In contrast to the results based on imputed data, the two plots are far more similar, with Holsteins and Friesians splitting into clusters along principal component 1 for both $\mathbf{G}_{50k}$ and $\mathbf{G}_{HD}$, with all HF cross animals clustering with the Friesians.

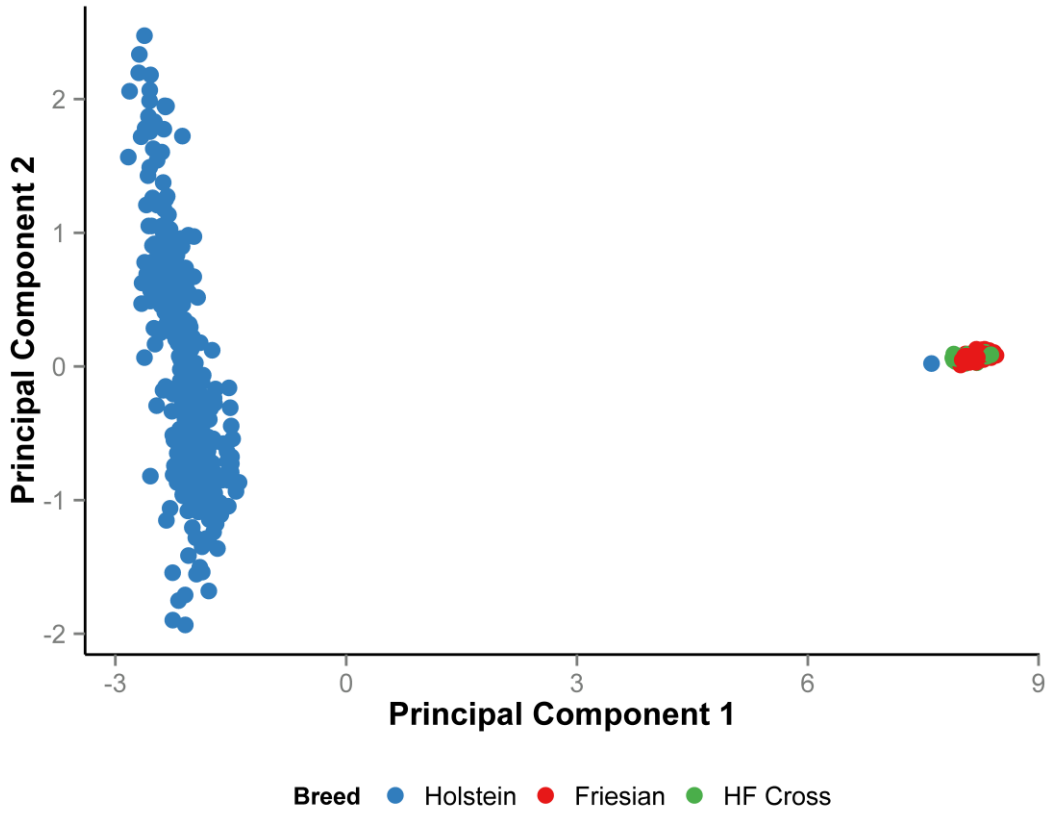

Figure 2.8 Principal components 1 and 2 based on a principal components analysis of the genomic relationship matrix $\mathbf{G}_{50k}$, with individuals coloured by breed.

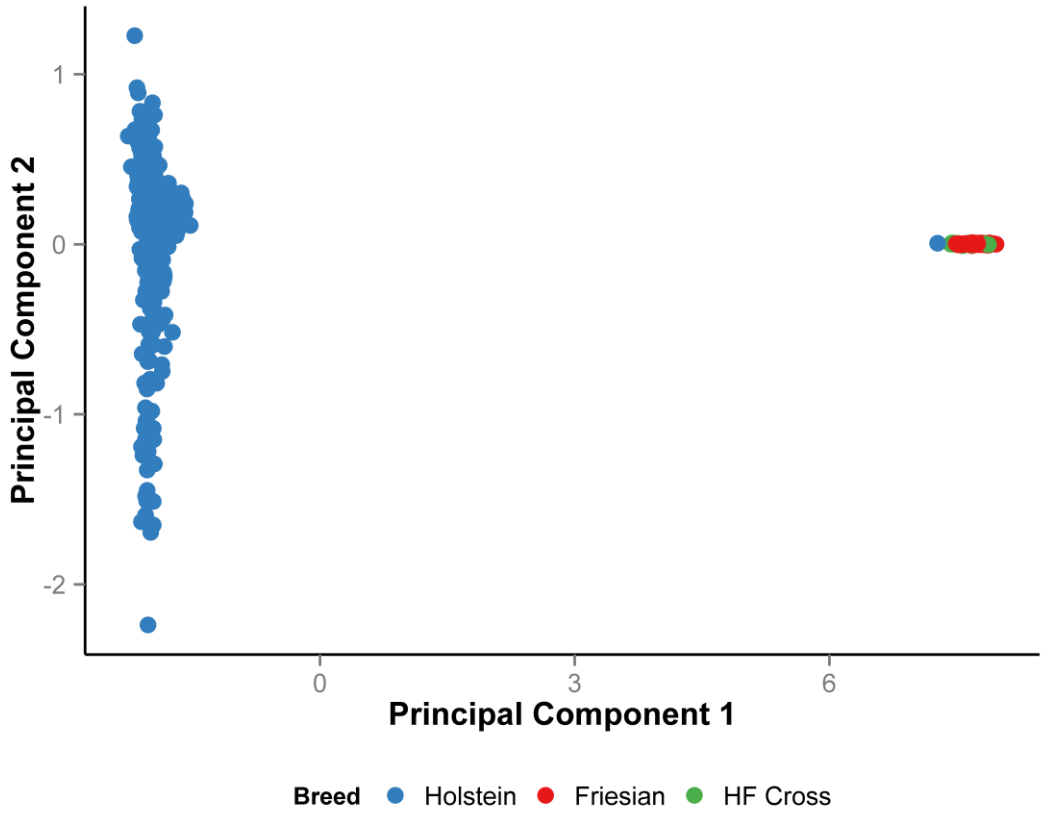

Figure 2.9 Principal components 1 and 2 based on a principal components analysis of the genomic relationship matrix $\mathbf{G}_{50k}$, with individuals coloured by breed.

2.3.3 LD decay

The mean r^2 between adjacent markers on the HD chip was 0.45 (s.d = 0.39), based on 536,198 SNP pairs. For markers common to the 50k chip, mean r^2 between adjacent markers was 0.04 (s.d = 0.15), based on 36,627 SNP pairs. Figure 2.8 shows the mean r^2 at a chromosome level for the HD and 50k chip, which ranged from 0.40 to 0.48 for the HD chip, and from 0.02 to 0.6 for markers common to the 50k chip. The mean distance between markers on the HD chip was 4.7kb, and 67.8kb between markers common to the 50k chip.

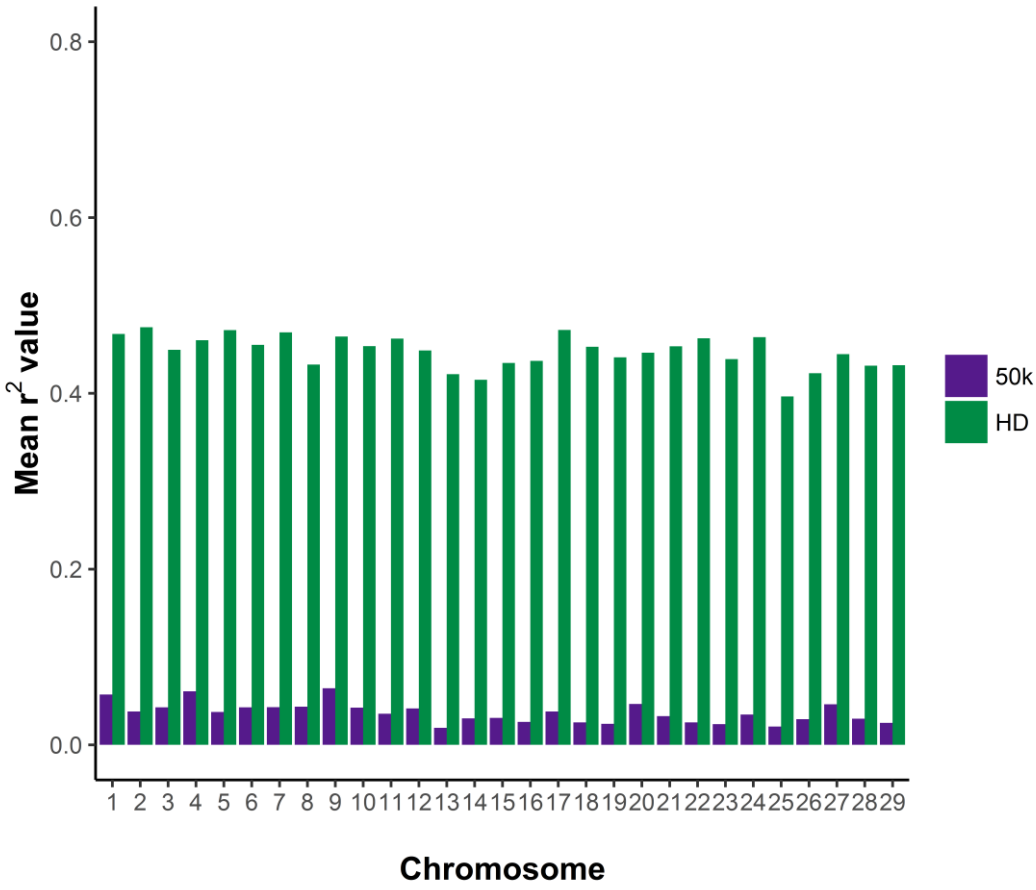

Figure 2.10 Mean LD per chromosome based on markers from the HD chip, and markers from the 50k chip. LD is expressed as the squared correlation between alleles across the full population.

Figure 2.11 shows the decay of LD across a 1Mb region for both the HD chip and the 50k chip. Mean LD for the overall population based on the HD chip decreased from 0.31 to 0.06 across the region. Mean LD for each of the pure breeds was slightly higher than observed in the overall population. When only markers common to the 50k chip were considered, mean LD for the overall population was much lower, starting at 0.05 and decreasing to 0.02 as the distance between markers increased. The mean LD for Holsteins based on 50k markers mirrored that of the overall population, whereas for Friesians it was higher, starting at 0.22 and decreasing to

0.05 as distance increased. Mean LD for Friesians based on 50k markers mirrored that of the HD markers from 0.1Mb onwards.

LD between markers was also calculated for non-imputed animals only (n = 413), to investigate whether imputation was responsible for the lower LD in Holsteins than in Friesians. The corresponding LD decay plot is shown in Figure 2.12. The level of LD in Friesians based on the 50k chip is again comparable to that seen with the HD chip. However, in this case, LD for the overall population based on the HD chip was lower for non-imputed animals than when imputed genotypes were used. Overall population LD was lower than the level of LD seen in Holsteins when using the HD chip, but higher than that seen in Holsteins when using only markers common to the 50k chip.

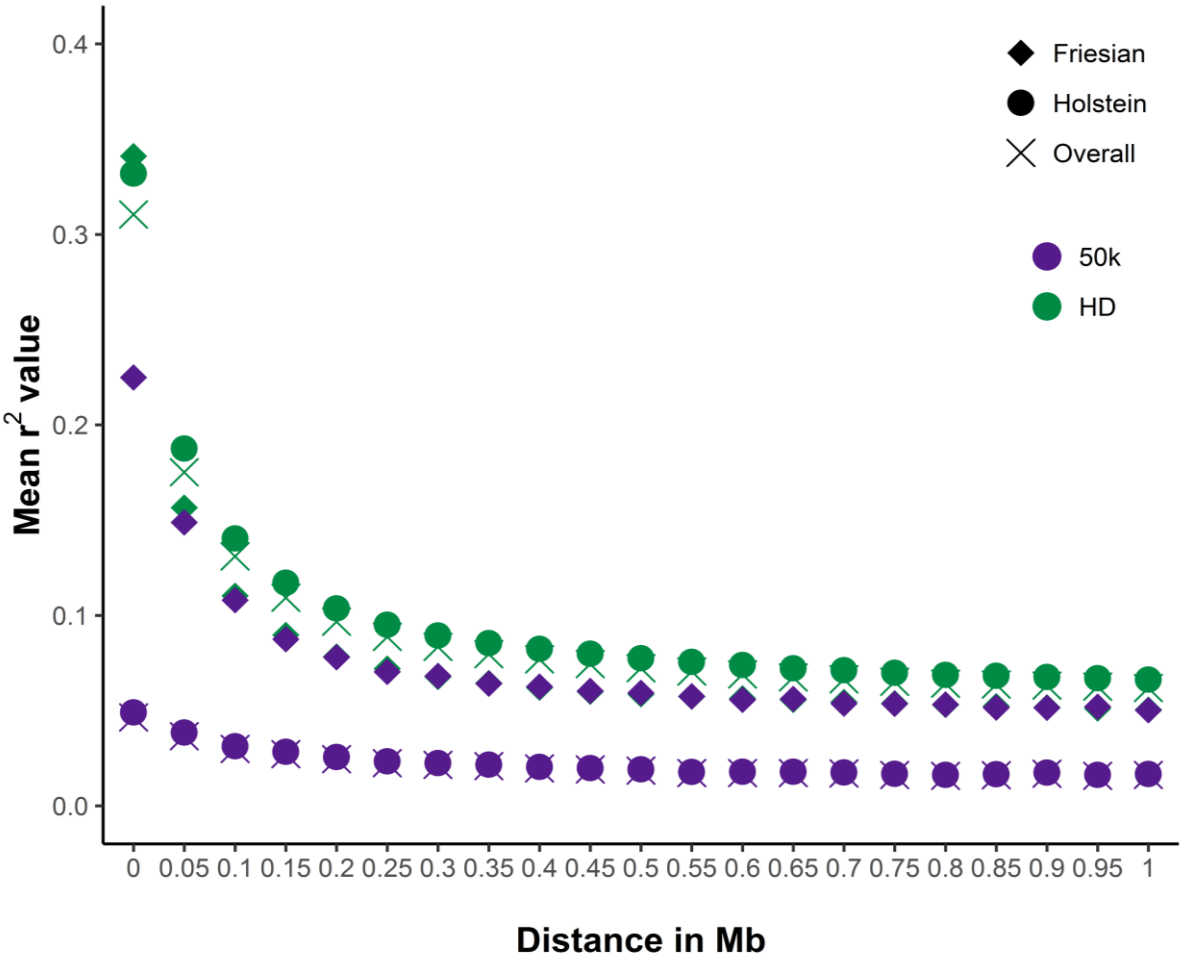

Figure 2.11 Mean LD for SNPs across a 1Mb region, based on SNPs from the HD chip, and also SNPs common to the 50k chip only. Mean LD (expressed as the squared correlation between alleles) is shown for the overall population and also for both pure breeds.

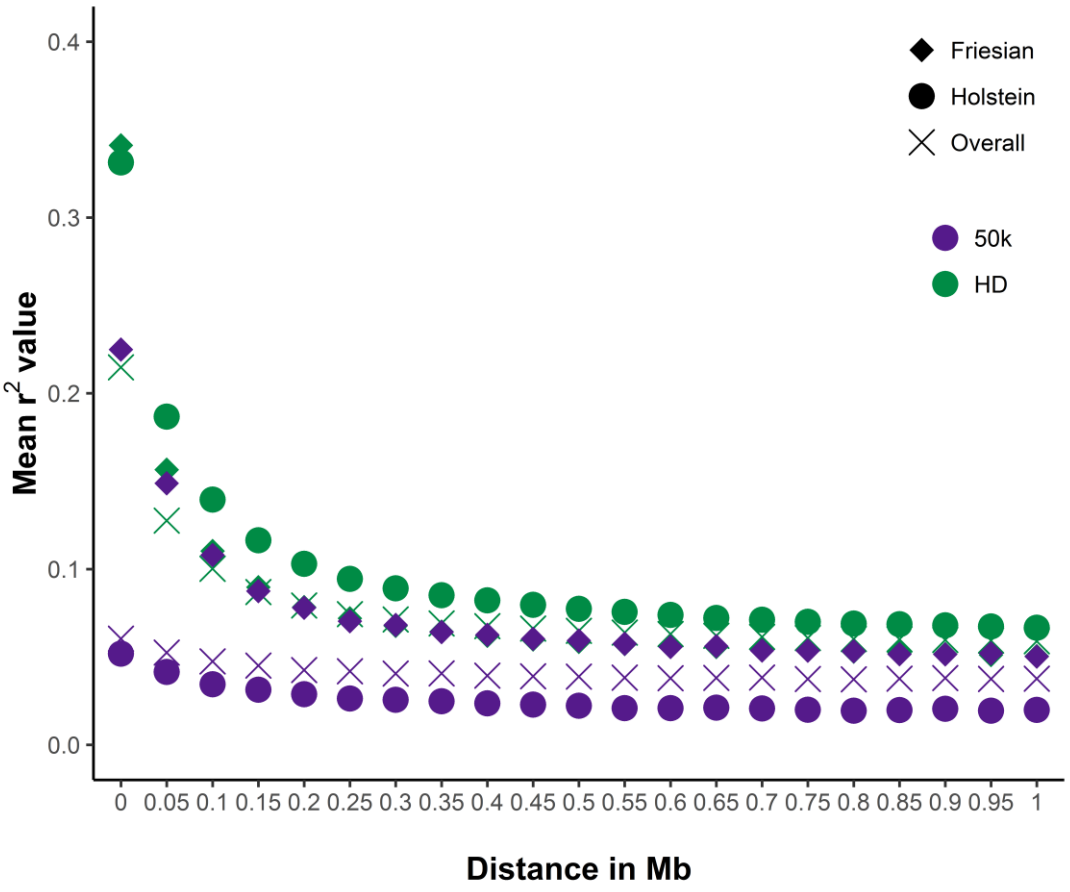

Figure 2.12 Mean LD for SNPs across a 1Mb region, based on SNPs from the HD chip, and also SNPs common to the 50k chip, for non-imputed genotypes only. Mean LD (expressed as the squared correlation between alleles) is shown for the overall population and also for both pure breeds.

2.3.4 Genomic evaluation accuracy

Accuracy of evaluation for all five traits can be seen in Figures 2.13 to 2.17 respectively, along with an estimate of the prediction bias. Accuracy of evaluation ranged from -0.61 to 0.84 depending on the trait of interest, validation population and reference population used. Accuracies for the Friesian validation population based on full and Friesian only reference populations are also shown in Table 2.5. As expected, the Holstein only reference population was unable to estimate accurate GEBVs in Friesians for any of the traits of interest, with the same trends seen when

trying to use a Friesian reference population to estimate GEBVs in Holsteins. The majority of results showed higher prediction accuracy for Friesians when using the multi-breed reference population as opposed to the Friesian only reference population. However, none of the differences in accuracy between multi-breed (Full) and Friesian reference populations proved to be statistically significant ($p = 0.11$ to $p = 0.98$). Overall, higher accuracies were observed when using the HBLUP method, and when using HD SNPs, with differences between chip types being more prominent when the HBLUP method was used. Higher accuracies of prediction were observed for the three production traits than for the non-production traits.

An element of prediction bias (regression coefficient $\neq 1$) was present in all analyses, but there was no consistency as to whether GEBVs were being over or under-estimated in comparison to dEBVs. The level of prediction bias was lower when using the HBLUP method for four of the five traits analysed. GEBVs calculated using SNPs from the 50k chip showed less bias than those using the HD chip in four out of five traits.

Due to previous analysis uncovering a potential issue with regards to using imputed data, GEBVs were also calculated using non-imputed data only. Results based on non-imputed genotypes mirrored those based on the full data set, with no statistically significant differences in accuracy observed between the dataset using imputed genotypes and the dataset using non-imputed genotypes only ($p = 0.64$ to $p = 1$).

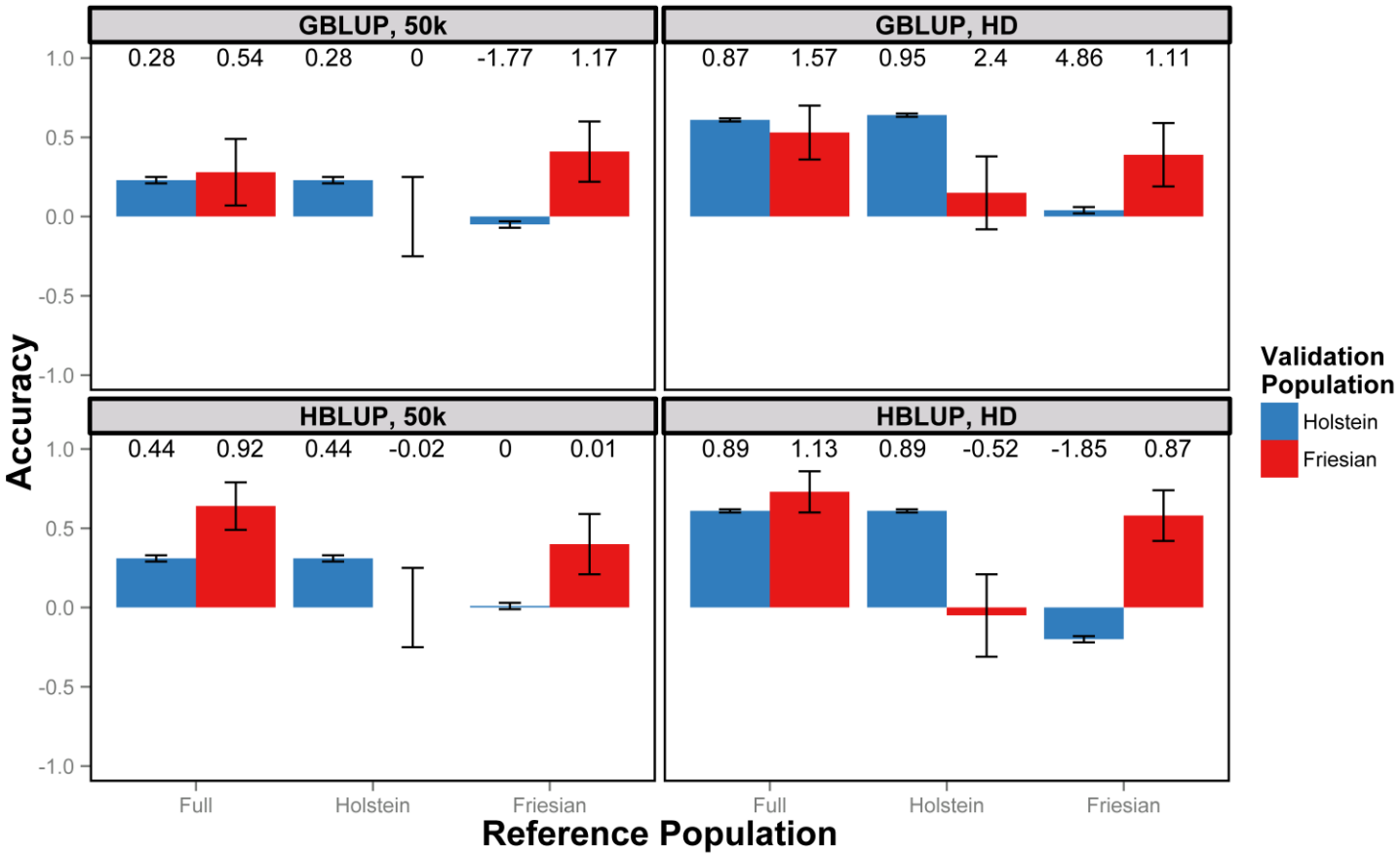

Figure 2.13 Accuracy of GEBVs for milk yield for Holstein and Friesian validation populations, calculated using two methods (GBLUP and HBLUP), using data from two chip types (HD, 50k) and based on three different reference populations (Full, Holstein only and Friesian only). Error bars show the standard error of accuracy, and the values along the top indicate the regression coefficient.

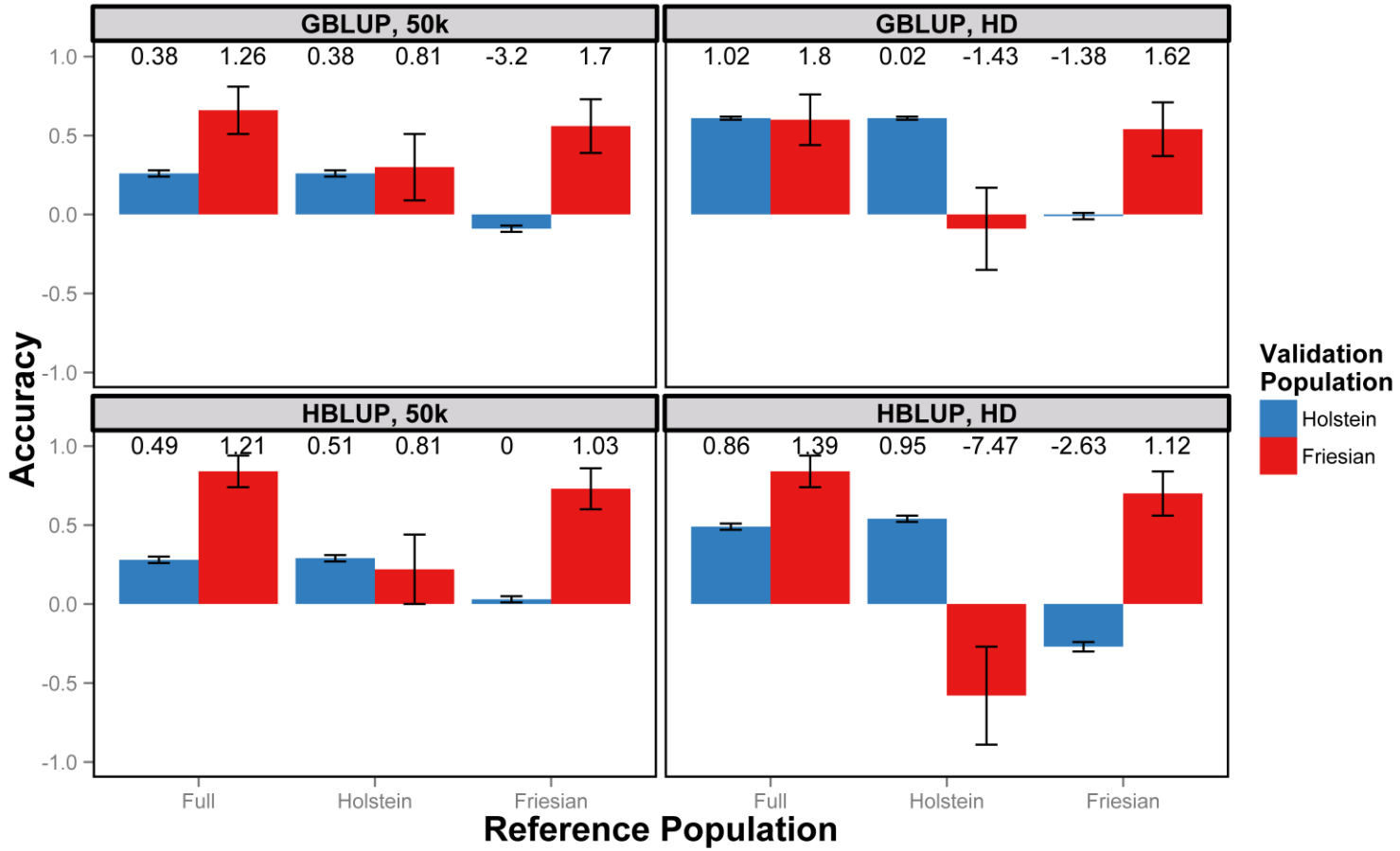

Figure 2.14 Accuracy of GEBVs for fat yield for Holstein and Friesian validation populations, calculated using two methods (GBLUP and HBLUP), using data from two chip types (HD, 50k) and based on three different reference populations (Full, Holstein only and Friesian only). Error bars show the standard error of accuracy, and the values along the top indicate the regression coefficient.

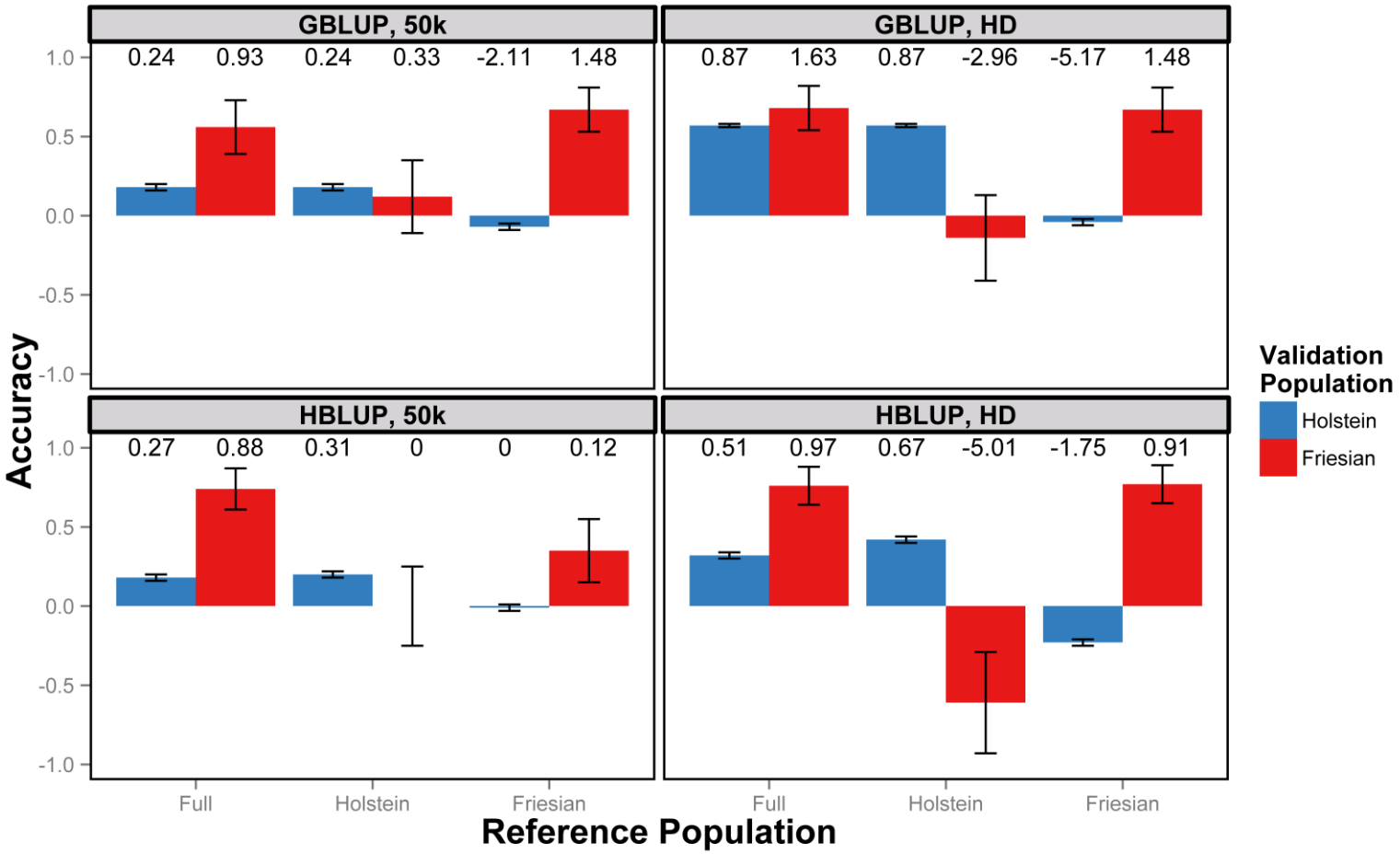

Figure 2.15 Accuracy of GEBVs for protein yield for Holstein and Friesian validation populations, calculated using two methods (GBLUP and HBLUP), using data from two chip types (HD, 50k) and based on three different reference populations (Full, Holstein only and Friesian only). Error bars show the standard error of accuracy, and the values along the top indicate the regression coefficient.

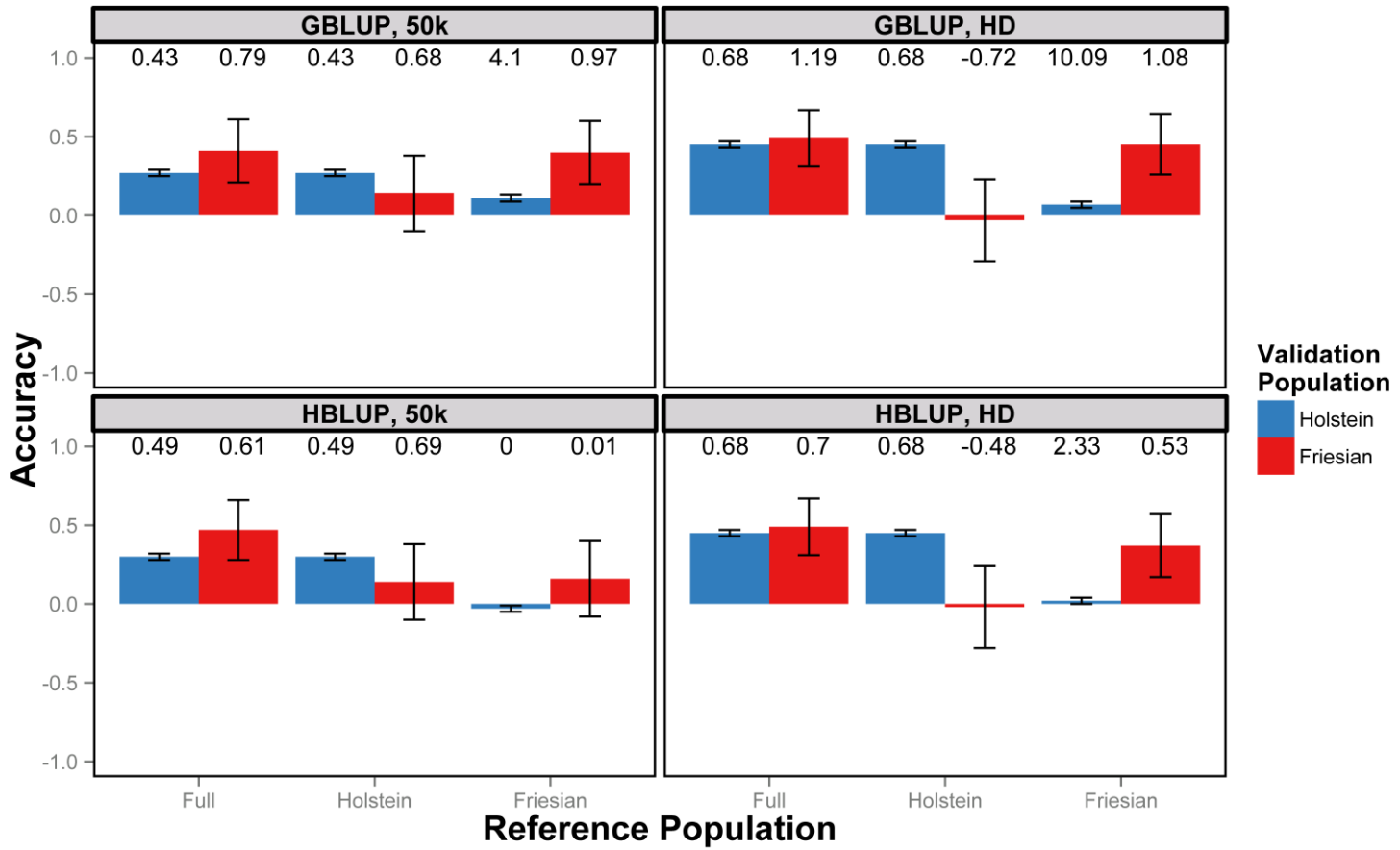

Figure 2.16 Accuracy of GEBVs for lifespan for Holstein and Friesian validation populations, calculated using two methods (GBLUP and HBLUP), using data from two chip types (HD, 50k) and based on three different reference populations (Full, Holstein only and Friesian only). Error bars show the standard error of accuracy, and the values along the top indicate the regression coefficient.

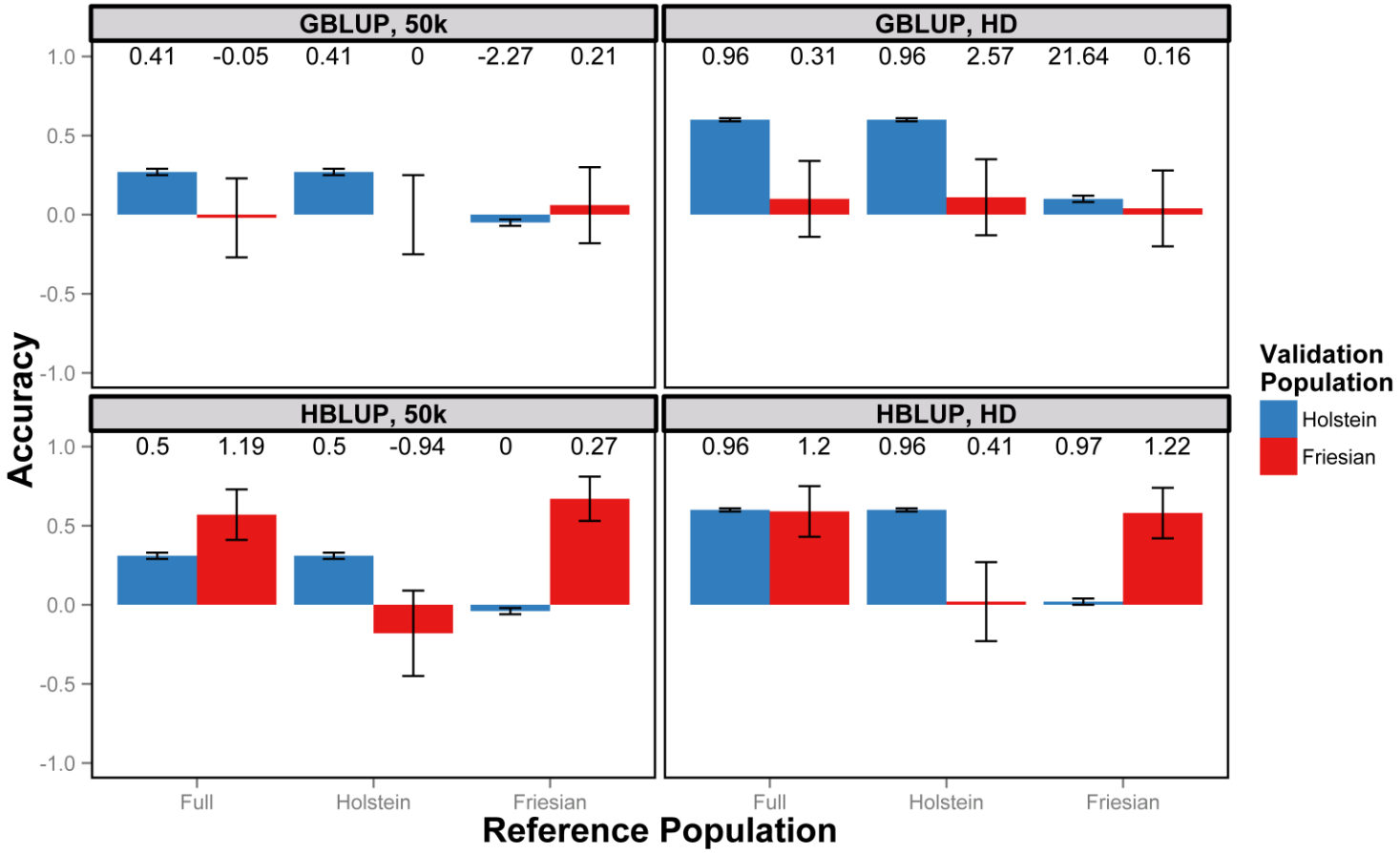

Figure 2.17 Accuracy of GEBVs for somatic cell count for Holstein and Friesian validation populations, calculated using two methods (GBLUP and HBLUP), using data from two chip types (HD, 50k) and based on three different reference populations (Full, Holstein only and Friesian only). Error bars show the standard error of accuracy, and the values along the top indicate the regression coefficient.

Table 2.5 Accuracy of evaluation for the Friesian validation population based on the full reference population (r_{FULL}), and the Friesian only reference population (r_{FRI}). The p values relate to the difference between r_{FULL} and r_{FRI} and have been calculated using the Fisher r to z transformation.

Trait	Method	Chip	r_{FRI} (s.e)	r_{FULL} (s.e)	p
Milk Yield	GBLUP	50k	0.41 (0.19)	0.28 (0.21)	0.68
		HD	0.39 (0.20)	0.53 (0.17)	0.62
	HBLUP	50k	0.40 (0.19)	0.64 (0.15)	0.36
		HD	0.58 (0.16)	0.73 (0.13)	0.47
Fat Yield	GBLUP	50k	0.56 (0.17)	0.66 (0.15)	0.66
		HD	0.54 (0.17)	0.60 (0.16)	0.81
	HBLUP	50k	0.73 (0.13)	0.84 (0.10)	0.42
		HD	0.70 (0.14)	0.84 (0.10)	0.33
Protein Yield	GBLUP	50k	0.67 (0.14)	0.56 (0.17)	0.62
		HD	0.67 (0.14)	0.68 (0.14)	0.96
	HBLUP	50k	0.35 (0.20)	0.74 (0.13)	0.11
		HD	0.77 (0.20)	0.76 (0.12)	0.94
Lifespan	GBLUP	50k	0.40 (0.20)	0.41 (0.20)	0.98
		HD	0.45 (0.19)	0.49 (0.18)	0.89
	HBLUP	50k	0.16 (0.24)	0.47 (0.19)	0.36
		HD	0.37 (0.20)	0.49 (0.18)	0.70
Somatic Cell Count	GBLUP	50k	0.06 (0.24)	-0.02 (0.25)	0.83
		HD	0.04 (0.24)	0.10 (0.24)	0.87
	HBLUP	50k	0.67 (0.14)	0.57 (0.16)	0.65
		HD	0.58 (0.16)	0.59 (0.16)	0.97

The highest accuracy for Friesian animals for each of the five traits of interest is shown in Table 2.6, along with the accuracy of parent average information. The Full reference population gave higher accuracies of evaluation for Friesians for four of the five traits of interest. All five traits were better predicted using HD genotype data and the HBLUP method of prediction, with the exception of lifespan, where GBLUP and HBLUP accuracies predicted using HD data and the Full reference population were equal. Accuracies obtained via genomic evaluation surpassed those based on parent average information for all five traits.

Table 2.6 The highest accuracy achieved for each trait for the Friesian validation population (r_{GEBV}), and the corresponding accuracy expected from parent average for each trait (r_{PA}). r_{PA} was calculated by AHDB Dairy based on records for calves currently alive.

Trait	Reference	Method	Chip	r_{GEBV}	r_{PA}
Milk Yield	Full	HBLUP	HD	0.73	0.58
Fat Yield	Full	HBLUP	HD	0.84	0.58
Protein Yield	Friesian	HBLUP	HD	0.77	0.58
Lifespan	Full	HBLUP	HD	0.49	0.47
Somatic Cell Count	Full	HBLUP	HD	0.67	0.52

2.4 Discussion

2.4.1 PCA and comparisons of G matrices.

PCA based on $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$ differed greatly when all genotypes were used to create G. It became clear that the discrepancies between the two G matrices was due to issues with imputed data, as when the analysis was repeated having removed imputed genotypes, the first two principal components were then comparable, with the first two principal components explaining similar percentages of observed variation. The results confirm that the Holstein and the British Friesian are in fact two genetically distinct breeds.

Principal components are commonly incorporated as fixed covariates in genome-wide association studies to account for population structure, reducing the potential for spurious associations. Inclusion of principal components into genomic evaluation models does not have the same effect. Daetwyler et al (2012) showed that including principal components decreases the accuracy of evaluation in a multi-breed sheep population, and hypothesised that population structure contributed more to prediction accuracy than the LD between markers and QTL. Janss et al (2012) explored the use of principal components in genomic selection, and described how principal components are accounted for in the random part of the model, and so to fit them as fixed covariates leads to the effects being "double counted". We therefore refrained from incorporating any data from the PCA into the final model.

When all elements of the **G** matrix are considered and correlated with each other, the correlation between the elements of $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$ are lower than we would expected

for relationships among Holstein animals, at 52% correlation as opposed to >90% for all other breed relationship combinations. Accuracy of imputation could be partly responsible for this, as when imputed genotypes were excluded from the G matrix calculations the correlation between $\mathbf{G_{50k}}$ and $\mathbf{G_{HD}}$ for Holstein/Holstein relationships rose from 0.52 to 0.71. This correlation was still significantly lower than for other breed relationship groups. However, the accuracy of imputation estimated by findhap averaged between 0.97 and 0.99, so these estimates are either inaccurate, or the lower correlation may be due to other factors.

Each of the G matrices contained both negative elements, and elements far greater than 1, which is not something that is generally observed in the numerator relationship matrix used for traditional genetic evaluations. However, the properties of G are such that the average value of the off diagonal elements is zero, and the average value of the diagonal elements is 1, and so for this to be the case some elements will be larger or smaller than expected. Simeone et al. (2011) also observed a multi-modal distribution of diagonal values of **G** in a broiler population, and concluded that the birds with a value of 1.5 or greater belonged to a separate line of broilers. It was suggested that the diagonal values of G could be used as a diagnostic tool to identify secondary populations in a data set (Simeone et al., 2011), and this is precisely what is observed here for our population of Holstein and Friesian animals.

A number of studies have surmised that the presence of familial relationships between individuals (i.e. genetic linkage) has a bigger impact on the accuracy of genomic selection than LD (Wientjes et al., 2013; Luan et al., 2012). We therefore

regarded the pedigree relationship matrix (**A**) constructed in this study as a benchmark to which genomic relationships may be compared (Figure 2.6 to 2.7, Table 2.4). The significantly higher correlation between $\mathbf{G_{HD}}$ and **A** than between $\mathbf{G_{50k}}$ and $\mathbf{G_{50k}}$ and **A** suggests that the HD markers are better able to estimate relationships between pairs of Holstein animals. This result is in concordance with both Goddard et al (2010) and Luan et al (2009), who note that relationships based on high density are likely to be a better estimator of relationships at the level of QTLs, albeit still not a perfect estimate.

2.4.2 LD decay

The extent of LD in this population is slightly lower than what has previously been observed in cattle populations (McKay et al., 2007). Mean LD for adjacent SNPs was approximately ten-fold higher for SNPs from the HD chip than those common to the 50k chip. However, the average LD for SNPs 50kb or less apart on the 50k chip is over two-fold higher for Friesian animals than for Holstein animals. We also see little difference between LD between the two chips at distances of 100kb or greater, suggesting lower genetic variation in the Friesian population than the Holstein population. The most likely explanation for this is small effective population size (Ne) in the Friesian population. When Ne is small, genetic drift is more likely to lead to loss of genetic variation via random genetic drift, and therefore higher levels of LD. Methods have been developed to estimate Ne from LD (Waples and Do, 2010), however Ne estimates can be highly biased when sample size is small (England et al., 2006) therefore we did not attempt to estimate Ne in this study. The difference in the persistence of LD in the two breeds based on the different chip sizes also matches

the results that we see in the correlation between the two G matrices, suggesting that it may be lower levels of LD in the Holstein population when using the 50k chip as opposed to the HD chip that lead to differences in relationship coefficients.

The low variation observed for this population of Friesian animals could also be a function of small sample size, however, bulls that are selected for genotyping are generally proven bulls that have a large influence on the wider population, and so we may expect that the level of LD observed within this population is indicative of the wider Friesian population.

The level of LD at 50k for Holstein animals is lower than seen in populations of Chinese and Nordic Holsteins (Zhou et al., 2013), and in a small population of US Holsteins (McKay et al., 2007). The Holsteins used in our study originate from a number of countries across Europe, as well as the USA. A previous study has indicated that Holsteins worldwide can be considered a homogenous population (Zenger et al., 2007), however the study in question was based on fewer than 1000 SNPs, which were not sampled evenly across the genome. It is possible that although we consider the Holsteins used in our study as a single population, they could differ at a genetic level due to differences in breeding schemes in the various countries of origin. In contrast, the Friesians used in our study are sampled from the UK and Ireland only, and this may contribute to why they appear to be a more homogeneous population than the Holsteins. This difference between the levels of LD based on the 50k and HD chips in our study could also explain the low correlation between Holstein genomic relationships from $\mathbf{G}_{\mathbf{50k}}$ and $\mathbf{G}_{\mathbf{HD}}$.

2.4.3 Genomic evaluation accuracy

It has been documented that to accurately predict genomic breeding values for a particular breed, that the breed in question should be represented in the reference population (Hayes et al., 2009a; Olson et al., 2012; Pryce et al., 2011). Zhou et al (2014b) were able to predict GEBVs for Danish Red cattle based on a Nordic Holstein reference population, however it was noted that the Danish Red had a high genomic relationship with the Nordic Holstein. The accuracies achieved in this study support the conclusions made in the Hayes, Olson and Pryce studies, as accuracies of predicting Friesian GEBVs from Holstein training data - and vice-versa - resulted in low (<0.10) or negative prediction accuracy. This reflects the results from the PCA on the genomic relationship matrix, which show the Holstein and British Friesian to be two genetically distinct populations. The large negative accuracies observed in some cases are most likely to have arisen due to chance considering the small sample size of the Friesian validation population.

2.4.4 Using a multi-breed reference population

The primary objective of this study was to assess whether it would be possible to improve the accuracy of Friesian GEBVs by incorporating Holstein data into the reference population, as opposed to using a small Friesian reference population. For four out of five traits, the highest accuracy was observed using the Full reference population, suggesting that incorporating Holstein genotypes as part of a multi-breed reference population is beneficial when calculating Friesian GEBVs. However, none of the increases in accuracy when using the Full reference population instead of the

Friesian reference population were statistically significant. This was due to the small number of Friesians in the validation population (n = 17 to n = 18). We were unable to increase the number of Friesians in the validation population as to do so would decrease the number of Friesians available in the reference population. This would have further lessened our ability to draw inferences between using a multi-breed versus a single-breed reference population for genomic predictions, and would likely have a negative impact on the observed prediction accuracy (Erbe et al., 2012; Hozé et al., 2013). Hozé et al (2013) also saw small but non-significant increases in prediction accuracy for Normande bulls when using a multi-breed reference population, with a larger benefit being observed when the number of Normande bulls in the training set was small. Despite not reaching statistical significance, considering the small number of Friesian genotypes available, and the effect of reference population size on the accuracy of genomic evaluation (Goddard and Hayes, 2009), we suggest that a using a multi-breed reference population containing Holstein and Friesian animals would be preferable to a Friesian only reference population for the prediction of Friesian GEBVs.

2.4.5 HBLUP vs GBLUP

Our second objective was to assess whether incorporating phenotype data into the genomic evaluation model via the HBLUP method would improve the accuracy of evaluation for Friesian individuals. A number of studies including Mucha et al. (2015) and Carillier et al. (2014), have reported higher accuracies of prediction in crossbred and multi-breed populations when using the HBLUP method compared to GBLUP. This study also showed clear benefits to using HBLUP to incorporate

further Friesian phenotypes into the evaluation, with the majority of evaluation accuracies from HBLUP analyses being higher for Friesians than the equivalent accuracies obtained via GBLUP. The largest gains were seen for somatic cell count, where the inclusion of approximately 1,300 Friesian phenotypes was responsible for increases of between 0.49 and 0.61 depending on the reference population and chip density, which was far more than expected. Conversely, using the HBLUP method did not result in any increase in accuracy for lifespan, despite the inclusion of almost 5,000 extra Friesian phenotypes. These two non-production traits have a lower heritability (h^2_{SCC} = 0.11 and h^2_{LS} = 0.06) than the three production traits (h^2 = 0.47 to h^2 = 0.55). The number of extra Friesian phenotypes available for the HBLUP analysis for lifespan and somatic cell count was also much fewer than for the production traits. Lower heritability traits generally need more phenotype data available to accurately estimate GEBVs (Villumsen et al., 2009), and so we believe that the smaller volume of data and the low number of Friesians in the validation population are responsible for this inconsistency in performance of HBLUP among the non-production traits.

2.4.6 The impact of chip density

The third objective of the study was to investigate whether using a higher density SNP chip would increase the accuracy of evaluation for Friesians. De Roos et al (2008) investigated the persistence of phase across Holstein-Friesian, Jersey, and Angus cattle, and concluded that using upwards of 300k SNP markers would allow the detection of LD from before breed divergence. In this study, with regards to the Friesians we observed an increase in accuracy for the majority of traits when the Full

reference population was used, but the benefit of the HD chip was less obvious when considering results from the Friesian only reference population. Other studies in cattle that have compared the utility of the HD chip with the 50k chip have seen little or no increase in the accuracy of evaluation when SNPs from the HD chip were used (Su et al., 2012; Erbe et al., 2012). Again, due to the small number of Friesians in the validation population, none of the increases in accuracy using the HD chip were significant, and this effect may just be an artefact of the data.

With regard to the Holsteins, the HD chip performed significantly better than the 50k chip ($p = 0$) using both the Full and the Holstein reference populations. Some of this difference may be due to issues with imputation as discussed above, however, significant differences between the two chip sizes remained for all traits except lifespan when only non-imputed genotypes were used in the analysis ($p = 0$ to $p = 0.08$).

The increase in accuracy when using the HD chip for Holstein genomic evaluation was higher than expected, as previous studies have not reported any significant increases in accuracy when moving from the 50k to the HD chip (Erbe et al., 2012; Ertl et al., 2014; Su et al., 2012). We believe that the increase in accuracy of Holstein GEBVs is due to the higher persistence of LD between HD SNPs than 50k. The study by Su et al. (2012), reported the average pair-wise LD between adjacent 50k SNPs in the Nordic Holstein population as 0.21, which was greater than the 0.04 calculated in our study.

2.4.7 Prediction bias

An estimate of the prediction bias was obtained as the slope of the regression, where a value of 1 signifies no bias. Regression coefficients ranged from -7.47 to 21.64, with some level of prediction bias present in the majority of evaluations. Extremely high and extremely low values were observed when one breed was used to predict the other, particularly in the low heritability traits. A higher level of bias was seen for low heritability traits compared to high heritability traits in the study by Luan et al. (2009), though the bias in that study was of a smaller magnitude. Luan et al. also suggest that the use of dEBVs in place of phenotypes may also be a source of bias, which could account for some of the bias we see in this study.

2.4.8 Scope for further work

Bayesian methods of prediction have also been used for multi-breed genomic evaluations (Su et al., 2012; Hozé et al., 2013; Erbe et al., 2012). Hayes et al. (2009b) suggest that Bayesian methods may perform better across breeds, when comparing GBLUP and Bayesian methods, both Erbe et al. (2012) and Su et al (2012) saw slightly higher accuracies of prediction when using Bayesian methods. We were unable to estimate GEBVs using a Bayesian prediction method due to time constraints, however we suggest that a method such as BayesC be should potentially be investigated before implementing Friesian evaluations on a commercial scale, but we hypothesise that the benefit of adding in a large number of extra Friesian genotypes via the HBLUP method will outweigh the benefit of a Bayesian model, at least for more polygenic traits. A higher accuracy may be observed with Bayesian

methods for traits such as fat yield, where known QTL have a high impact on the trait (Luan et al., 2009). The large negative accuracies seen in this study for some traits when one breed is used to predict another, could be due to SNPs being in opposite linkage phases in the two breeds (Riedelsheimer et al., 2013), or simply have occurred by chance due to the small sample size of the Friesian validation population.

Another aspect to be considered in future studies is whether the trait of interest should be considered as a single trait across the two breeds, or rather as two distinct but correlated traits (a "multi-trait" model). We originally carried out an analysis analysis for production traits as two correlated traits, but as a result of the genotyping errors mentioned previously, time did not allow for this step to be carried out with the current data set. Previous studies have shown that in a multi-breed scenario, multi-trait models have resulted in slightly higher accuracy of prediction (Olson et al., 2012; Makgahlela et al., 2013), and deserve further attention (Lund et al., 2014). An alternative approach could be to implement the multi-compartment model suggested by Hamidi-Hay and Rekaya (2014), which would allow the effect of a SNP marker to differ between breeds. If the ultimate aim were to implement a single multi-breed commercial genomic evaluation for all UK bulls, this model may be of use.

2.4.9 Conclusions

Although the Holstein and British Friesian breeds are considered to be closely related, there are clear differences between the two breeds at the genetic level, and

one cannot simply be used to predict the other. Although the limited number of Friesian genotypes available makes it difficult to draw completely robust conclusion, we suggest that combining Holstein and Friesian genotypes into a multi-breed reference population can facilitate the estimation of Friesian GEBVs. The incorporation of more phenotypes via the HBLUP method can be used to maximise accuracy, and in practice should be used. The utility of using HD genotypes is less clear, but the results merit further investigation. Uptake of the HD SNP chip in commercial situations has been limited due to the cost per genotype, with breeders placing more importance on number of genotypes over genotype density. This issue could be bypassed by obtaining the majority of HD genotypes via imputation rather than direct genotyping, however care must be taken to ensure that genotypes are accurately imputed.

Chapter 3: Genomic selection in a crossbred cattle population using data from the Dairy Genetics Project for East Africa

3.1 Introduction

"Across-breed" genomic evaluations fall into two broad categories, multi-breed evaluations, where two or more breeds are combined into a single reference population to allow genomic evaluation of numerically small breeds, and crossbred evaluations, where the aim is to predict the performance of crossbred individuals. The previous chapter explored the multi-breed scenario, and this chapter explores the crossbred scenario, using exclusively crossbred animals in the analysis. The body of this chapter has been published as a short communication in the Journal of Dairy Science.

Contributions to the paper were as follows; J Ojango, M Okeyo and J Gibson provided data for analysis. R Mrode carried out calculation of yield deviations. Genomic analyses were completed by myself and R Mrode, and the manuscript was composed by myself. R Mrode and M Coffey assisted in experimental design and provided comments on the manuscript.

3.2 Short communication

Short communication: Genomic selection in a crossbred cattle population using data from the Dairy Genetics Project for East Africa

A. Brown,[*] J. Ojango,[†] J. Gibson[+], M. Coffey,[*] M. Okeyo,[†] R Mrode[*†1]

* Animal & Veterinary Sciences

Scotland's Rural College

Easter Bush, Midlothian EH25 9RG, Scotland, United Kingdom

† International Livestock Research Institute (ILRI)

Box 30709, Nairobi, Kenya

[+]University of New England

Armidale, NSW 2351, Australia

[1] Corresponding author: raphael.mrode@sruc.ac.uk

ABSTRACT

Due to the absence of accurate pedigree information, it has not been possible to implement genetic evaluations for crossbred cattle in African small-holder systems. Genomic selection techniques that do not rely on pedigree information could, therefore, be a useful alternative. The objective of this study was to examine the feasibility of using genomic selection techniques in a crossbred cattle population using data from Kenya provided by the Dairy Genetics East Africa Project. Genomic estimated breeding values for milk yield were estimated using 2 prediction methods, GBLUP and BayesC, and accuracies were calculated as the correlation between yield deviations and genomic breeding values included in the estimation process, mimicking the situation for young bulls. The accuracy of evaluation ranged from 0.28 to 0.41, depending on the validation population and prediction method used. No significant differences were found in accuracy between the 2 prediction methods. The results suggest that there is potential for implementing genomic selection for young bulls in crossbred small-holder cattle populations, and targeted genotyping and phenotyping should be pursued to facilitate this.

SHORT COMMUNICATION

Genomic selection is now widely used in the dairy industry; with genomic estimated breeding values (GEBV) now being commercially produced for several breeds worldwide, as part of routine genetic evaluations. However, the majority of these evaluation schemes are carried out in developed countries, where most animals evaluated are purebred, and have large volumes of phenotype, genotype, and pedigree data. In developing countries, such as those in Eastern Africa, a large proportion of dairy production is carried out by small holders, who in many cases keep fewer than 10 cattle. These cattle are mostly crosses between indigenous African breeds and exotic dairy breeds, and have little phenotypic or pedigree data available. It has, therefore, not been possible to implement conventional genetic evaluation methods in these populations. As a result, bulls cannot currently be effectively ranked for genetic progress, preventing effective genetic improvement. If the level of phenotypic recording can be increased, and sufficient funding is available to cover the costs of genotyping, genomic selection may be a suitable tool for estimation of breeding values in these crossbred cattle.

Several studies have highlighted the potential for crossbred genomic evaluations using a training population made up of crossbred animals (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Mucha et al., 2015; VanRaden and Cooper, 2015). Earlier studies, such as those by Ibáñez-Escriche et al. (2009) and Toosi et al. (2010), used simulated data to investigate the potential for using a crossbred reference population to estimate breeding values of purebred animals for the performance of their crossbred offspring. Results suggested that there is potential for using crossbred

reference populations to predict GEBV in purebreds, with no necessity to use complex models to assign breed-specific allele frequencies. More recently, VanRaden and Cooper (2015) used empirical data to show that genomic-predicted transmitting abilities can be computed for crossbred animals by applying purebred marker effects that have been weighted by the crossbred animal's genomic breed composition. In a study involving UK dairy goats, Mucha et al. (2015) computed milk yield GEBV for crossbred goats using a crossbred training population. The results suggested that there was no additional benefit to using SNP-BLUP to estimate breeding values, compared with pedigree-based BLUP, but higher accuracies were achieved when the single step method was implemented.

The above studies used a range of statistical methods for prediction of GEBV, with Ibánez-Escriche et al (2009) and Toosi et al. (2010) using Bayesian methods of prediction, whereas Mucha et al. (2015) implemented SNP-BLUP and single step approaches. Simulation studies have suggested that Bayesian methods have a slight advantage over GBLUP methods for genomic prediction (Hayes et al., 2009b); however, the methods have not been compared using real-world data in the analysis of dairy traits.

This study aims to investigate the feasibility of using genomic selection in a small population of African crossbred cattle, using 2 statistical methods, GBLUP and BayesC. The method of assessing achieved accuracy mimics the situation of young bulls.

The data set consisted of genotype data for 1,013 cows aged 4 to 8 years, from the Kenyan component of the Dairy Genetics East Africa Project (Ojango et al., 2014, Gibson et al. 2014) Animals consisted of varying crosses between indigenous African breeds (N'dama–*Bos taurus*, and Nellore–*Bos indicus*) and 5 exotic dairy breeds (Ayrshire, Friesian, Holstein, Guernsey, Jersey). All individuals were genotyped using the Illumina BovineHD BeadChip (Illumina, San Diego, CA). Genotype data were edited by loci; SNP with a minor allele frequency of <0.05, a call rate of <0.95, or with no chromosomal position, were removed, along with those that were detected as not being in Hardy-Weinberg equilibrium and SNP on the X chromosome. After applying these filters, 665,408 autosomal SNP were available for analysis.

The phenotypes used were milk yield deviations (YD). These were computed from a fixed test-day model using test-day records for the first 3 lactations with management group, year-month of test, parity, and dairy group by breed interaction fitted as fixed effects. In addition, fixed lactation curves of Legendre polynomials of order 4, nested within dairy group by breed interaction, were fitted to account for crossbreeding effects in the model (J. Ojango, unpublished data). Random effects of animal and permanent environment were also included in the model. The YD were averaged by cow and the corresponding weight for YD for each cow used in the genomic analysis was computed as the inverse of the standard error. The heritability of milk yield based on this model was 0.30.

A genomic relationship matrix was computed for all animals using VanRaden's first definition of **G** (VanRaden, 2008), where $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$, where **Z** is a design matrix of centred genotypes, and p_i is the allele frequency estimated across breeds for the major allele at SNP *i*. Principal components analysis (PCA) was then carried out on the **G** matrix using the R function "princomp" (R Core Team, 2013), to investigate the genomic relationships between individuals.

Figure 3.1 illustrates the results of the PCA. Although there is no distinct separation between clusters of animals, the first principal component divides the animals into 5 groups based on the proportion of their genetics that is contributed by exotic dairy breeds, their so-called percentage exotic breeds. Due to this clustering, we chose to investigate how well GEBV for animals with the highest and lowest percentage exotic breeds could be estimated using the remainder of the population. Two groups with higher percentage exotic breeds were chosen for validation: (1) animals with percentage exotic breeds above 87.5%, and (2) animals with 60 to 87.5% exotic breeds. However, the number of animals with a low percentage of exotic breeds was too low to create a third validation population based purely on this category. The data were therefore re-organized into 6 categories, with each category defined by the combination of exotic breeds that contributed most of the exotic genes to the cross. These categories were (a) Ayrshires; (b) Friesians; (c) Ayrshires and Friesians; (d) Guernseys and Friesians; (e) Ayrshires, Friesians, and Guernsey; and (f) mixed exotic. For animals in category f, the exotic genes came from more than 3 exotic breeds (average percentage exotic breeds was approximately 46%), with indigenous

breeds contributing ≥40% of genetics in most cows. To represent animals with mainly indigenous genetics, a third validation group was created using animals from category f. Figure 3.2 shows the same PCA with animals labelled according to the 6 categories described above. Summary statistics for the 3 validation groups are shown in Table 3.1.

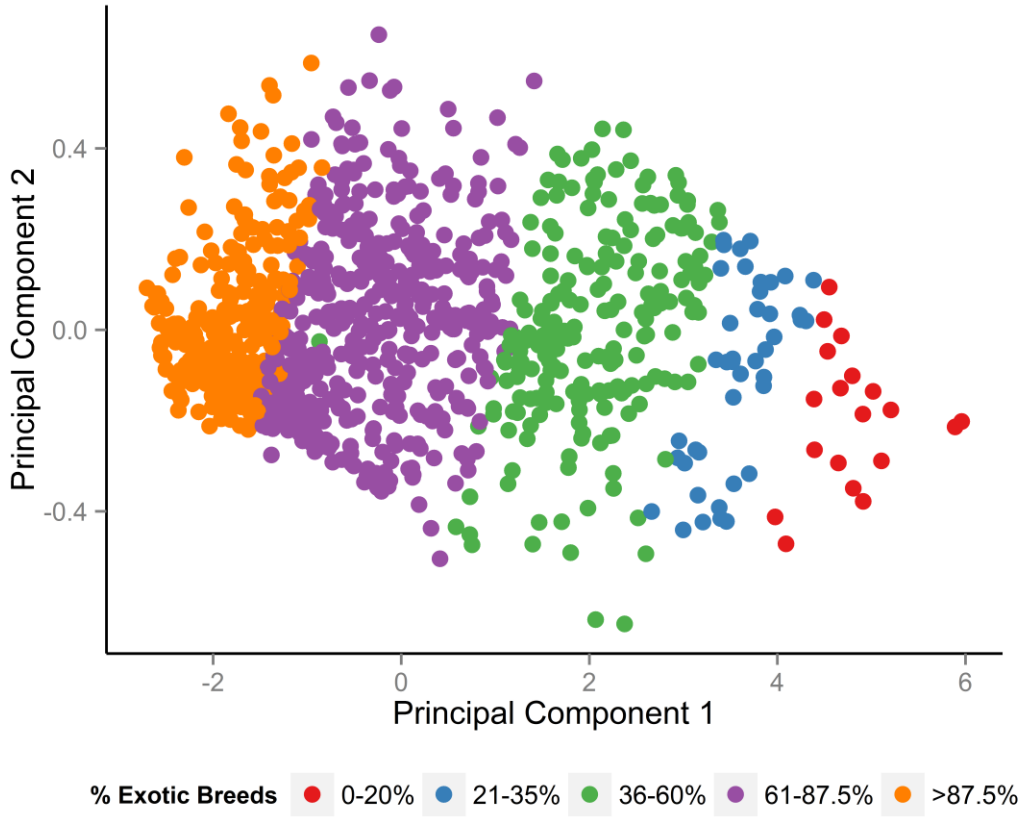

Figure 3.1 Principal components 1 and 2 based on the analysis of the genomic relationship matrix of 1,013 crossbred cows. Animals are labelled according to the percentage of their genetics contributed by exotic dairy breeds.

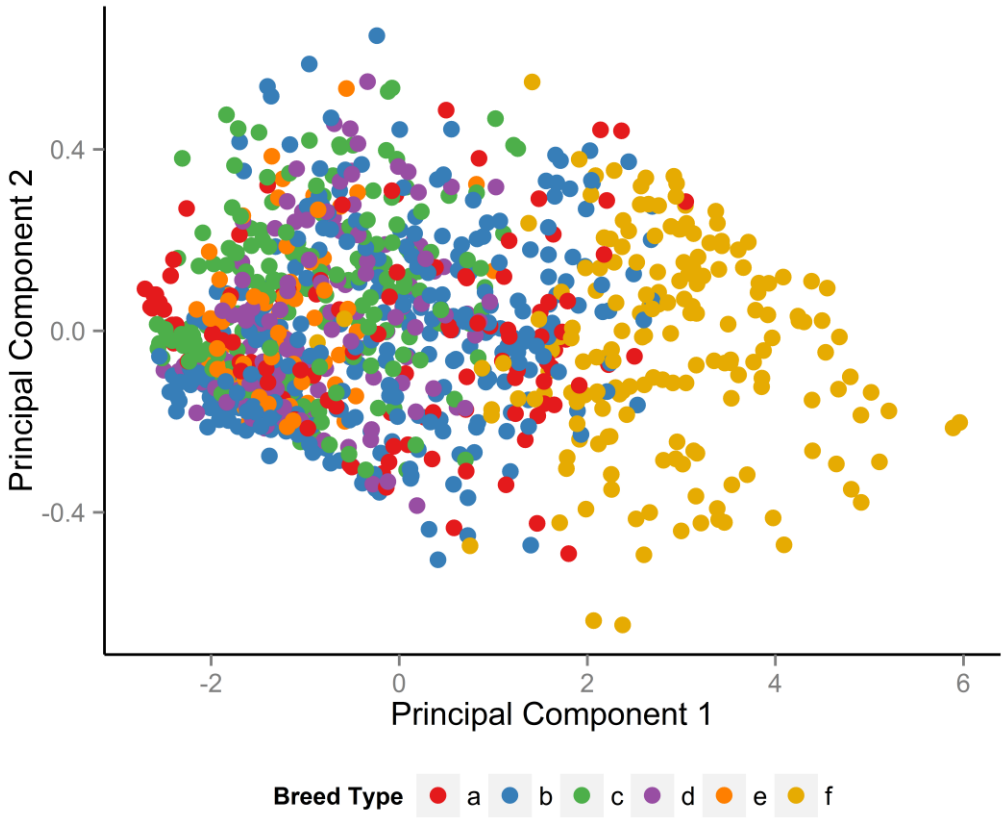

Figure 3.2 Principal components 1 and 2 based on the analysis of the genomic relationship matrix of 1,013 crossbred cows. Animals are split into 6 categories, with each category defined by the number of exotic breeds that contributed most of the exotic genes to the cross. a) Ayrshires, b) Friesians, c) Ayrshires and Friesians, d) Guernseys and Friesians e) Ayrshires, Friesians and Guernsey and f) Mixed exotic.

Table 3.1 Summary statistics for each of the three groups chosen for GEBV estimation and validation.

Validation group	Description	N	Mean yield deviation (s.d)	Range
1	>87.5% exotic	297	0.39 (1.60)	-2.34 – 7.75
2	61-87.5% exotic	448	0.00 (1.34)	-3.41 – 7.32
3	33-50% exotic	178	-0.61 (1.08)	-2.87 – 3.45

Two statistical models were used to compare their performance, GBLUP and BayesC. The model for the GBLUP analysis was $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}$ is the vector of the weighted YD, $\boldsymbol{\mu}$ is the overall mean, $\mathbf{Z}$ is an incidence matrix relating individuals to records, $\mathbf{g}$ is a vector of random animal effects with an assumed

distribution of $N(0,\sigma^2_g\mathbf{G})$, where σ^2_g is the additive genomic variance and **G** is the genomic relationship matrix calculated as detailed above, and **e** is a vector of residual effects with an assumed distribution of $N(0,\sigma^2_e\mathbf{I})$, where σ^2_e is the residual variance and **I** is an identity matrix. The software package Mix99 (Lidauer and Strandén, 1999) was used for GBLUP analysis.

The BayesC method used the same basic model as detailed above, but in this case vector **g** is defined as $\sum_{i=1}^{N}(z_i a_i I_i)$, where z_i is the genotype at SNP *i*, a_i is the effect of SNP*i*, and I_i is an indicator variable which is set to 1 if the *i*th SNP has an effect on the trait of interest, or 0 if it has no effect. The distribution of a_i is assumed $N(0, \sigma^2_a)$, where σ^2_a is the SNP variance. The SNP effects were assumed to be normally distributed, and variable *I* was assumed to be binomially distributed with probability π. Previous analyses using a BayesCπ model suggested a value of π of 0.23, and so in this study the value of π was set to 0.3. A custom written Fortran program following the method by (Mrode, 2014) was used for BayesC analysis.

The accuracy of prediction was calculated for both methods as the correlation between the YD and the GEBV within each of the 3 validation populations described above. In each case, the reference population comprised all animals in the data set that were not chosen for the validation. Reference and validation population sizes for each analysis are shown in Table 3.2

Table 3.2 Accuracies of GEBV based on GBLUP and BayesC models, for each of three validation groups; 1) animals with percentage exotic breeds above 87.5%, 2) animals with 60 - 87.5% exotic breeds, and 3) animals with predominantly indigenous genetics.

Validation Group	$N_{validation}$	$N_{reference}$	Accuracy (s.e) GBLUP	BayesC
1	297	716	0.41 (0.04)	0.39 (0.05)
2	448	565	0.35 (0.04)	0.35 (0.04)
3	178	835	0.32 (0.06)	0.28 (0.06)

Accuracies of prediction ranged from 0.28 to 0.41 dependent upon the validation group and the statistical method used (Table 3.2). The highest accuracies were observed for animals with percentage exotic breeds of above 87.5% (group 1), and lower accuracies for the mixed exotic group (group 3). In general, the validation accuracies reported for milk yield are much lower than observed in developed countries (Hayes et al., 2009b). Differences in size and type of data could be considered as major factors in this difference. However, this analysis provides the first estimates of genetic merit for this population and is, therefore, valuable for identifying extreme animals and selecting teams of young bulls that can be used for breeding. Small sample size in group 3 may be a factor contributing to a lower accuracy, as a small number of badly performing individuals can have a large effect on the overall accuracy. However, differences between accuracies achieved in the 3 validation groups were tested for significance using Fisher's r to z transformation, with no significant difference in accuracy between the 3 groups observed for either method of prediction ($P = 0.19$ to $P = 0.70$). Fisher's r to z transformation was also used to test the comparative performance of GBLUP and BayesC; no significant differences were found in performance between the 2 methods ($P = 0.68$ to $P = 1$).

It was particularly interesting that the BayesC method did not perform significantly better than the GBLUP model, as previous studies have suggested that Bayesian models should predict genomic breeding values with a higher accuracy than GBLUP (Hayes et al., 2009b). Bayesian methods of prediction require more computational time and greater computational power to run than GBLUP-based methods. Due to this difference in running time, GBLUP methods are often preferred in commercial situations; Bayesian methods must, therefore, produce substantially higher accuracies of prediction than GBLUP for the increased computational time to be worthwhile. As such, we suggest that the GBLUP model is more suitable for commercial evaluations of polygenic traits, such as milk yield, in crossbred populations. However, considering that Bayesian methods of prediction are expected to perform better for traits controlled by a small number of genes of large effect (Hayes et al., 2009b), we suggest that Bayesian models should still be considered when implementing evaluations for less polygenic traits.

The accuracies obtained in this study are similar to those reported by Mucha et al. (2015), who estimated GEBV for milk yield in a UK population of dairy goats. In the study by Mucha et al. (2015), the SNP-BLUP model did not outperform pedigree-based BLUP, and to see any benefit of implementing genomic selection, the authors had to incorporate further data using the single step method. We are unable to implement pedigree-based evaluation methods in this population of cattle; as such, we are comparing our predictions to a baseline accuracy of zero. The results presented above are, therefore, extremely positive, and provide an opportunity for undertaking selection and consequently increasing the rate of genetic progress within

this population. This study used high-density genotypes to capture as much genetic variation as possible within this crossbred population; however, it is unlikely that genomic selection will be implemented commercially using this chip due to the costs associated with high density genotyping. Work is currently on-going to develop a lower density chip that is suitable for use in the wider African small holder cattle population. As indicated earlier, the prediction of genomic merit in this study provides an opportunity for the selection of teams of young bulls for breeding, and will also help to identify extreme animals. It therefore provides the incentive for more targeted recording schemes that will allow the collection of more phenotypic data, with the aim of improving the accuracy achieved by increasing the size of the reference population. Innovative ways of giving timely and targeted feedbacks to farmers, based on such data, would help to support data collection and should be pursued.

ACKNOWLEDGEMENTS

The authors thank the Bill and Melinda Gates Foundation for funding the Dairy Genetics East Africa (DGEA) project. A. Brown also acknowledges the Biotechnology and Biological Sciences Research Council (Swindon, UK) and the Knowledge Transfer Network (London, UK) for funding.

3.3 Conclusion

This study demonstrates that there is potential for applying genomic evaluation techniques in crossbred cattle populations, but, as in the previous chapter, more data is needed to validate the work.

It has been noted that further clarification was necessary regarding the computation of genetic parameters for milk yield in this chapter, due to the lack of pedigree in the African crossbred dairy population. Due to this lack of pedigree information, genetic parameters were estimated with a genomic relationship matrix as opposed to a pedigree based numerator relationship matrix. It should also be documented that Nellore cattle are not indigenous to Africa as stated in the published paper, but are in fact indigenous to India.

As in chapter 2, this chapter has focussed on using high-density genotype data for genomic evaluations. Whole-genome sequence data is now being generated for multiple breeds of cattle, and so the final experimental chapter will focus on whether there is an advantage to using SNPs extracted from sequence data for across-breed genomic evaluation.

Chapter 4: Utilising sequence data to estimate across breed genomic relationships

4.1 Introduction

Within-breed genomic evaluations are out of reach of many dairy breeds due to small genotyped population size (Thomasen et al., 2014). Results from chapter 2 suggest that using HD genotypes that reduce the distance between SNP markers may improve the accuracy of a multi-breed genomic prediction, but this effect has not been mirrored in other studies, where only small increases have been observed both within and across breeds (Su et al., 2012; Erbe et al., 2012; Ertl et al., 2014).

It has been suggested that using whole-genome sequence (WGS) data would increase evaluation accuracy as it differs from genotype data, in that the causal variants (SNPs or other variants such as copy number variants or insertions/deletions) will likely be present in the data, eliminating the reliance on LD between markers and causal loci (Druet et al., 2013). Large volumes of WGS data have been hard to access due to the associated costs of sequencing, and so previous work relating to the use of WGS data for genomic prediction has either been based on simulation studies (Iheshiulor et al., 2016; Meuwissen and Goddard, 2010; Druet et al., 2013) or imputed data (van Binsbergen et al., 2015; Hayes et al., 2014). However, due to the combination of data-sharing consortiums such as the 1000 Bulls Project (Daetwyler et al., 2014) and the price per sequence falling as the technology evolves, implementing genomic evaluations based on WGS data is becoming increasingly viable.

Although results of genomic evaluations using WGS data within breeds have been mixed, Iheshiulor et al. (2016) reported an increase in prediction accuracy using a

combination of WGS data and a multi-breed reference population, suggesting WGS data may be of use in the quest to implement across breed genomic evaluations.

However, the volume of data derived from WGS is so much greater than that derived from genotyping chips, that the computational demand to process the data and run subsequent evaluations increases dramatically. Although WGS provides us with causal variants, the vast majority of mutations in a genome are neutral, and therefore unlikely to inform predictions. It is therefore not necessarily suitable to use WGS data in a commercial situation where evaluation methods are required not only to be robust but also regular, fast and efficient. The question arises as to whether it is possible to extract only the most informative elements of WGS data in order to facilitate its use in routine genomic evaluations. SNP variant predictors may provide a means of achieving this.

SNP variant predictors compare WGS data to a reference genome and can be used to annotate every variant in a dataset based on its location in the genome and its putative impact with regards to protein structure and behaviour. Annotation of WGS data using SNP variant predictors could allow the identification of SNPs that have a "significant" impact on the genome, allowing us to create a "custom or virtual" SNP chip. SNP selection for currently available chips is based on whether they are segregating across breeds (Matukumalli et al., 2009), whereas a custom chip based on WGS data from multiple breeds would allow segregating SNPs to be selected based on putative impact.

The present study aims to investigate the utility of SNPs extracted from the SNP variant predictor "snpEff" (Cingolani et al., 2012) in calculating genomic relationships between animals from three breeds, using data from 96 sequenced bulls provided by the Gene2Farm consortium (www.gene2farm.eu). While the number of individuals with sequence data is too small to facilitate the estimation of GEBVs, groups of novel SNPs will be assessed by calculating the correlation between relationships estimated using novel SNPs and relationships estimated using a high density panel of markers.

4.2 Methods

4.2.1 Data

Sequence data was made available for 96 bulls across three breeds (49 Brown Swiss, 31 Fleckvieh, and 16 Simmental). The Brown Swiss breed originated from the Alpine region of Switzerland in the 5th century, before spreading across the Alps and into Germany. The Simmental also originates from Switzerland, where it has traditionally been a dual-purpose breed. The Fleckvieh was developed by crossing Simmental cattle with local Bavarian breeds from what is now Austria and Germany, and is also considered to be a dual-purpose breed.

Sequence data was provided in variant call file format (vcf), with variants having been detected using the software freeBayes (Garrison and Marth, 2012). The vcf file contained 23,821,524 variants, of which 666,178 were common to the Illumina BovineHD genotyping chip, which contains 777,692 SNPs in total.

4.2.2 Variant annotation and filtering

The file was annotated using snpEff (Cingolani et al., 2012), with SNPs separated into four categories based on their putative impact on protein structure/behaviour. Number of SNPs in each category can be seen in Table 4.1.

SNPs annotated as having a high, moderate, or low putative impact on the genome were extracted for analysis, along with any SNPs common to the HD chip. All other modifier SNPs (n = 22,786,886) were excluded from the analysis. Table 4.2 shows the numbers of SNPs present in each category before quality control procedures were carried out.

Table 4.1 The number of variants relating to each impact category from snpEff. Impact descriptions are as described in the snpEff documentation (Cingolani, 2012)

Putative Impact	Impact description	No. SNPs
High	The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.	2,887
Moderate	A non-disruptive variant that might change protein effectiveness.	82,223
Low	Assumed to be mostly harmless or unlikely to change protein behavior.	103,194
Modifier	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.	23,633,220

Table 4.2 Number of SNPs in each category before quality control, where N_{novel} relates to SNPs that have been discovered from sequence data, and N_{HD} relates to SNPs that are present on the Illumina BovineHD SNP chip.

Category	N_{novel}	N_{HD}	Total
High impact	2,853	34	2,887
Moderate impact	79,837	2,386	82,223
Low impact	97,466	5,728	103,194
BovineHD chip	0	666,178	666,178

The resulting data was edited by individual animal and locus. Individuals with a call rate of <0.85 or that clustered with another breed in PCA were removed, leaving a data set containing data from 83 individuals (43 Brown Swiss, 25 Fleckvieh and 15 Simmental). The threshold call rate was reduced to 0.85 due to a large number of animals (44) having a call rate below the original threshold of 0.95. SNPs with a call rate of <0.95 were filtered out, along with those that were detected as not being in Hardy-Weinberg equilibrium and SNPs located on the X chromosome, leaving a dataset containing 544,288 autosomal SNPs. The inclusion of rare variants has been shown to have a positive impact on accuracy of genomic evaluations (Suchocki et al., 2014), and so to avoid removing rare variants from the dataset that potentially have a high influence, no filtering was carried out based on minor allele frequency (MAF). The number of SNPs in each impact category for the data set after quality control procedures is given in Table 4.3.

Table 4.3 Number of SNPs in each impact category post quality control

Category	No. SNPs
High	1,016
Moderate	27,975
Low	33,345
BovineHD	485,289
Total	544,288

4.2.3 Calculation of genomic relationship matrices and PCA

Genomic relationship (**G**) matrices were calculated for the following SNP combinations; High (H), High and Moderate (HM), High Moderate and Low (HML), HD, and all SNPs (ALL) in Table 4.3. The **G** matrices were calculated using two methods proposed by VanRaden (VanRaden, 2008), which differ in how they

incorporate allele frequency information. The first method (VR1) uses the following equation, $\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}'}{2\sum p_i(1-p_i)}$, where $\boldsymbol{Z}$ is a design matrix of centred genotypes, and p_i is the allele frequency estimated across breeds for the major allele at SNP i. The second method (VR2) uses the following equation, $\mathbf{G} = \mathbf{ZDZ}'$, where $\mathbf{Z}$ is a design matrix of centred genotypes, $\mathbf{D}$ is a diagonal matrix, where $\mathbf{D}_{ii} = m[2p_i(1-p_i)]$, where m is the number of SNPs used. All $\mathbf{G}$ matrices were calculated using the program preGSf90 (Misztal et al., 2002). A full list of the $\mathbf{G}$ matrices calculated is given in Table 4.4.

Principal component analysis was then performed on each of these matrices using the R function "princomp"(R Core Team, 2013), and the resulting principal components plotted to determine sources of variation between individuals.

Table 4.4 Full table of G matrices calculated for analysis, where H relates to High impact SNPs, HM relates to High and Moderate impact SNPs, HML relates to High, Moderate and Low impact SNPs, HD relates to SNPs common to the Illumina BovineHD SNP chip, and ALL relates to all SNPs in a data set. VR1 and VR2 relate to the method used to create **G**, with VR1 being VanRaden's first method, and VR2 being VanRaden's second method.

Category	VR1	VR2
H	G_{H_VR1}	G_{H_VR2}
HM	G_{HM_VR1}	G_{HM_VR2}
HML	G_{HML_VR1}	G_{HML_VR2}
HD	G_{HD_VR1}	G_{HD_VR2}
ALL	G_{ALL_VR1}	G_{ALL_VR2}

4.2.4 Comparison of G matrices

G matrices were compared by calculating the correlation between the relationship coefficients obtained from the off diagonals of different matrices. The assumption

was made that the **G** matrix calculated using ALL SNPs would give the most informative relationship information, and so the off diagonal elements of **G** from ALL would be regressed on the off diagonal elements of **G** for the other categories to provide an estimate of bias. In each case, correlations were calculated for each relationship combination separately. The different relationship combinations are detailed in Figure 4.1. Differences between correlations were tested for significance using the Fisher r-to-z transformation.

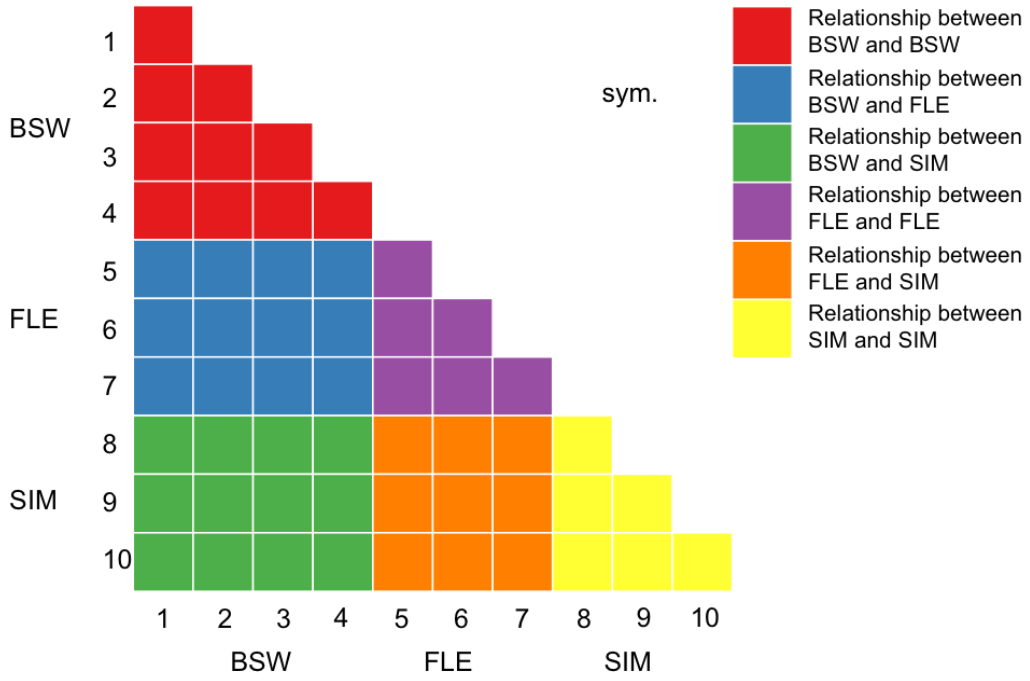

Figure 4.1 An example of the different relationships present within the genomic relationship matrix, where BSW refers to Brown Swiss, FLE refers to Fleckvieh, and SIM refers to Simmental individuals.

4.3 Results

4.3.1 Allele frequency and allele sharing across breeds

For SNPs to be informative across breeds, alleles must be segregating in all breeds of interest. To investigate the level of variation within each of the SNP categories, the minor allele frequency was calculated for each SNP within each of the three breeds. Figure 4.1 shows the spread of MAF for all three breeds for each category of SNPs. A higher proportion of SNPs with allele frequencies below 0.05 was observed for SNPs derived from the sequence data as opposed to those derived from the HD chip, regardless of impact category. The number of SNPs with a MAF of below 0.05 was approximately three times higher for novel SNPs (in all three categories) than observed in the HD SNPs.

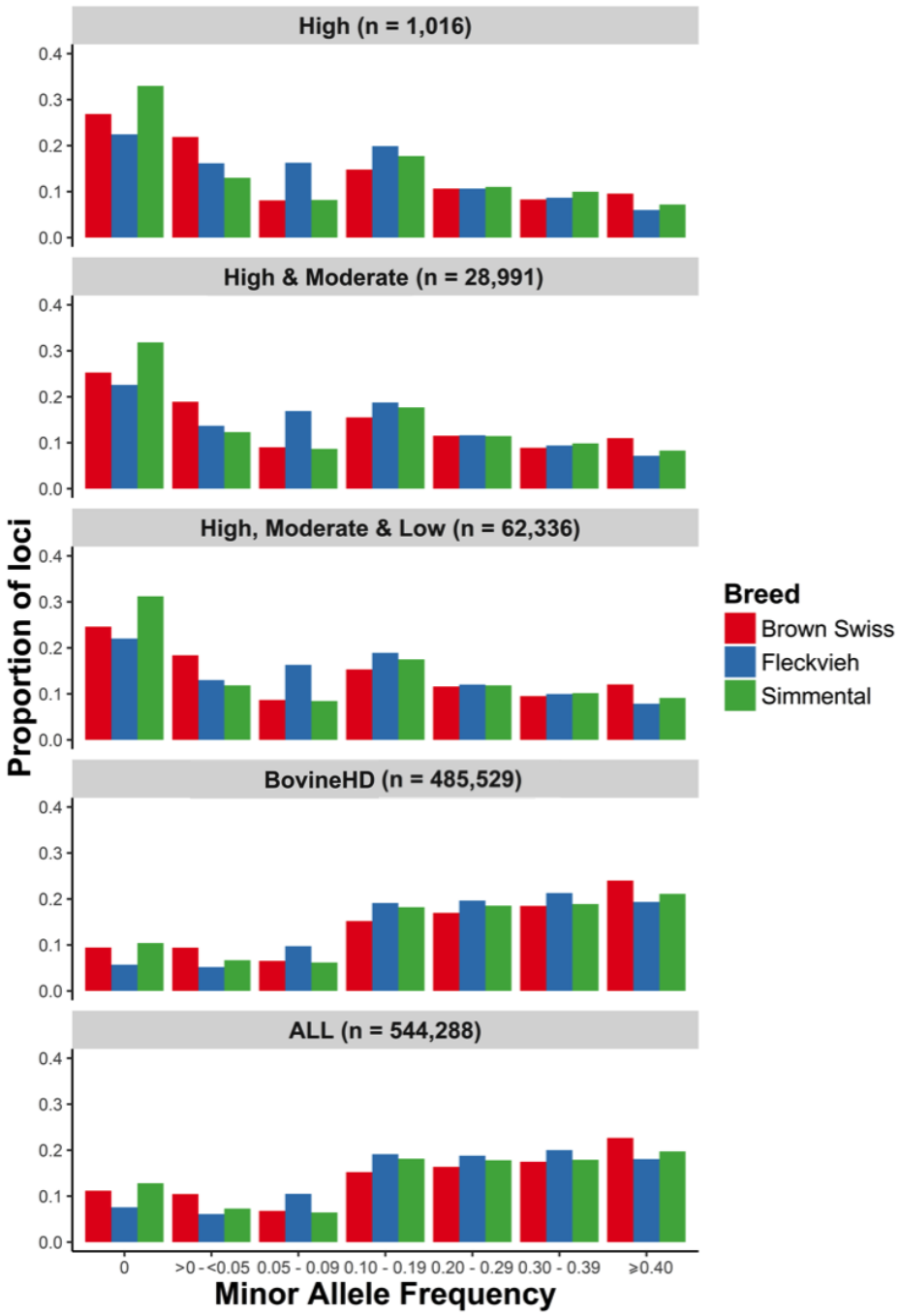

Figure 4.2 Proportion of loci at different minor allele frequencies for High, High & Moderate, High Moderate & Low, HD and ALL SNPs, separated by breed.

Though the minor allele frequency gives an indication of whether or not SNPs are segregating in each of the three breeds, it is not known whether the minor allele for a particular SNP is the same across the three breeds. The frequency of the reference

allele (i.e. the allele present on the UMD 3.1 *Bos taurus* reference genome) was therefore calculated for each SNP within each breed. SNPs were then ordered by the reference allele frequency observed in Brown Swiss, followed by the allele frequency in Fleckvieh and finally the reference allele frequency in Simmental, to compare reference allele frequencies across breeds. For ease of plotting, SNPs were split into 50 bins and the mean reference allele frequency for each breed was calculated for each SNP bin. Figures 4.2 to 4.4 compare the mean reference allele frequencies in Brown Swiss and Fleckvieh, Brown Swiss and Simmental, and Fleckvieh and Simmental respectively, and Table 4.5 shows the mean number of SNPs per bin for each category. For all categories, more variability was seen in the Fleckvieh and Simmental individuals than in the Brown Swiss, with fewer SNP bins fixed at the reference allele for these two breeds. The difference in mean reference allele frequency was smaller between Fleckvieh and Simmental than either of these breeds with the Brown Swiss, which was expected due to the fact that Fleckvieh and Simmental are more closely related. This is also illustrated by a higher correlation between reference allele frequencies in Fleckvieh and Simmental (Table 4.6).

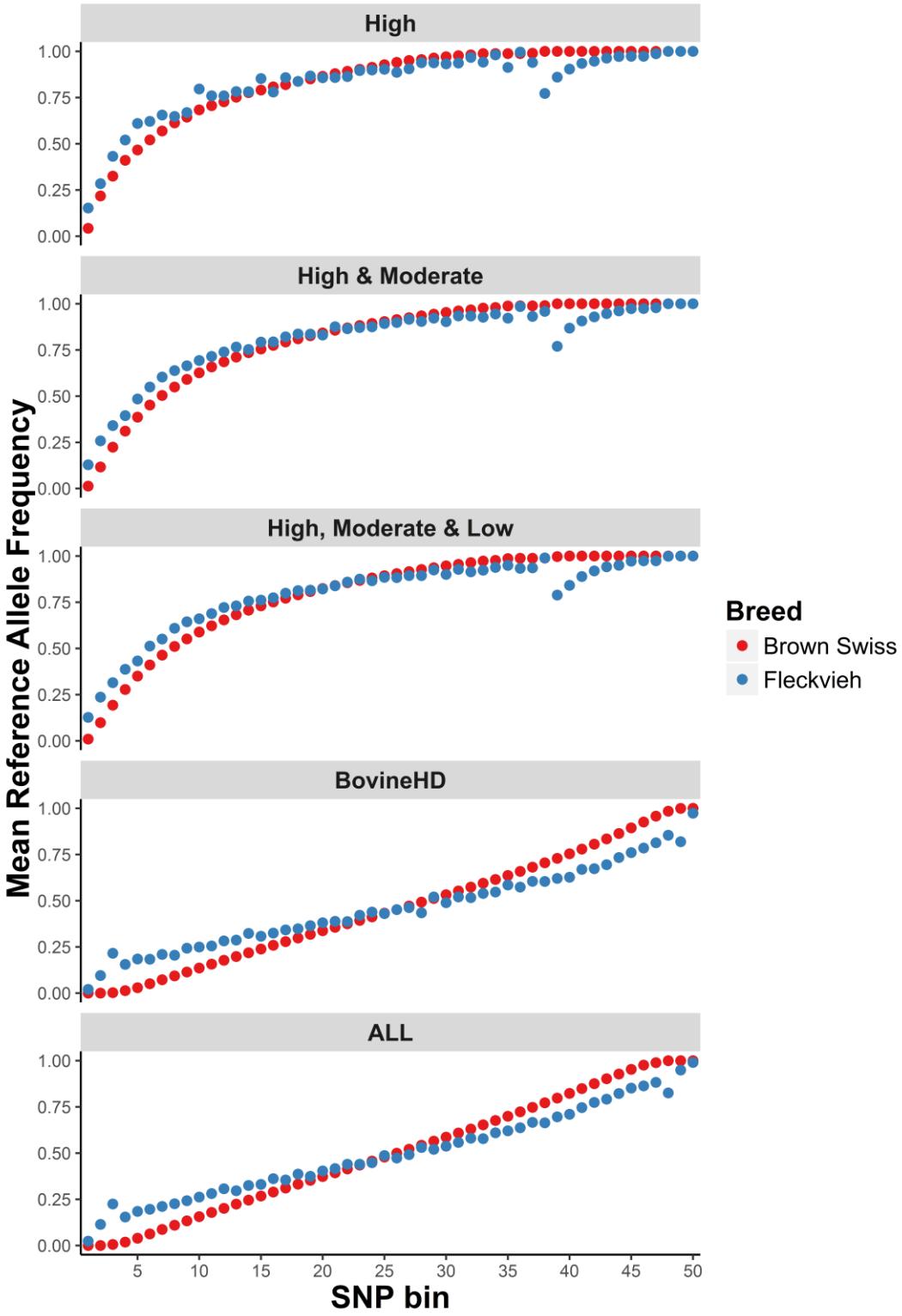

Figure 4.3 Mean reference allele frequency for Brown Swiss and Fleckvieh animals across 50 SNP bins.

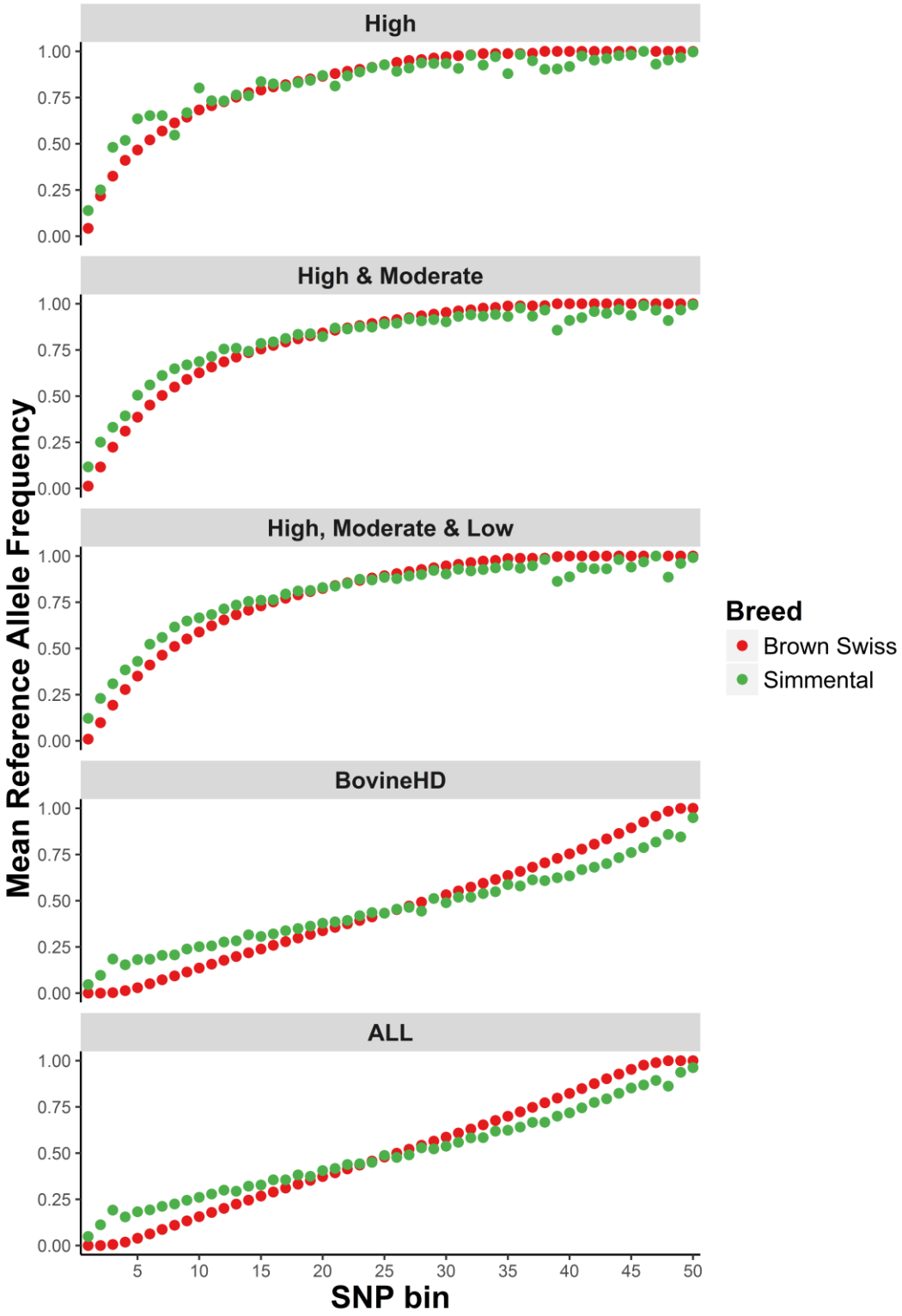

Figure 4.4 Mean reference allele frequency for Brown Swiss and Simmental animals across 50 SNP bins.

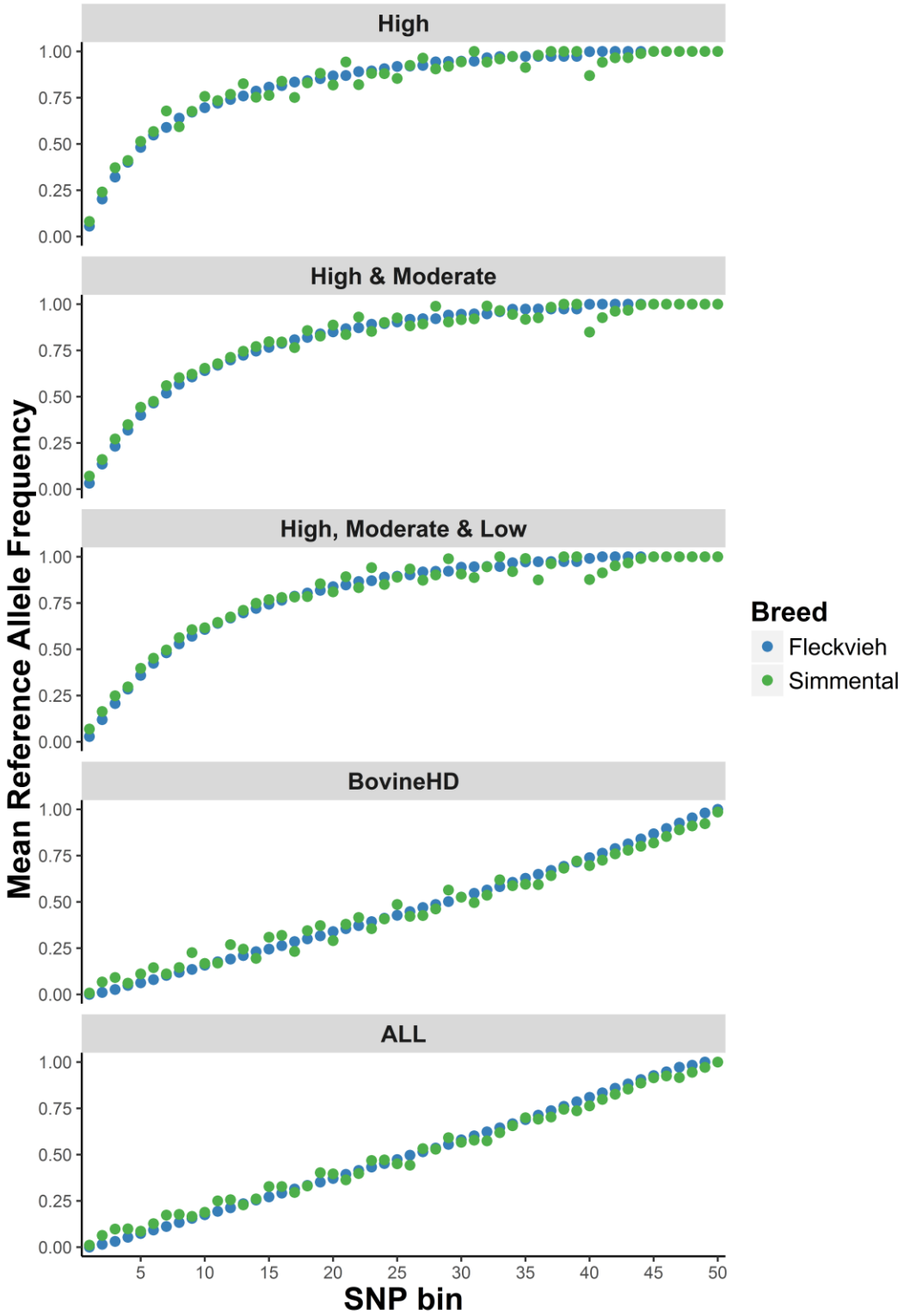

Figure 4.5 Mean reference allele frequency for Fleckvieh and Simmental animals across 50 SNP bins.

Table 4.5 Mean number of SNPs per bin (to the nearest SNP) for calculation of mean reference allele frequencies.

SNP category	Total SNPs	Mean SNP per bin
High	1,016	20
High & Moderate	28,991	580
High, Moderate & Low	62,336	1,247
BovineHD	485,529	9,711
ALL	544,288	10,886

Table 4.6 Correlation of reference allele frequency between breeds for ALL SNPs

Breed combination	r	s.e
BSW/FLE	0.77	0.0004
BSW/SIM	0.74	0.0004
FLE/SIM	0.87	0.0003

Venn diagrams were also plotted for High, Moderate, Low and BovineHD SNPs respectively, to see how many SNPs with MAF >0.05 were shared across breeds (Figures 4.6 to 4.8). Fleckvieh showed more variability than either Brown Swiss or Simmental across all four categories, with a higher number of SNPs with a MAF >0.05. Between 71.1% and 72.3% of SNPs were segregating in two or more breeds for each of the novel SNP categories (High, Moderate or Low), whereas 90.7% of SNPs from the BovineHD chip segregated across two or more breeds. Of the SNPs segregating in two breeds, a higher percentage of SNPs had MAF >0.05 in both Fleckvieh and Simmental breeds, than in Brown Swiss and Fleckvieh, or Brown Swiss and Simmental for all four SNP categories.

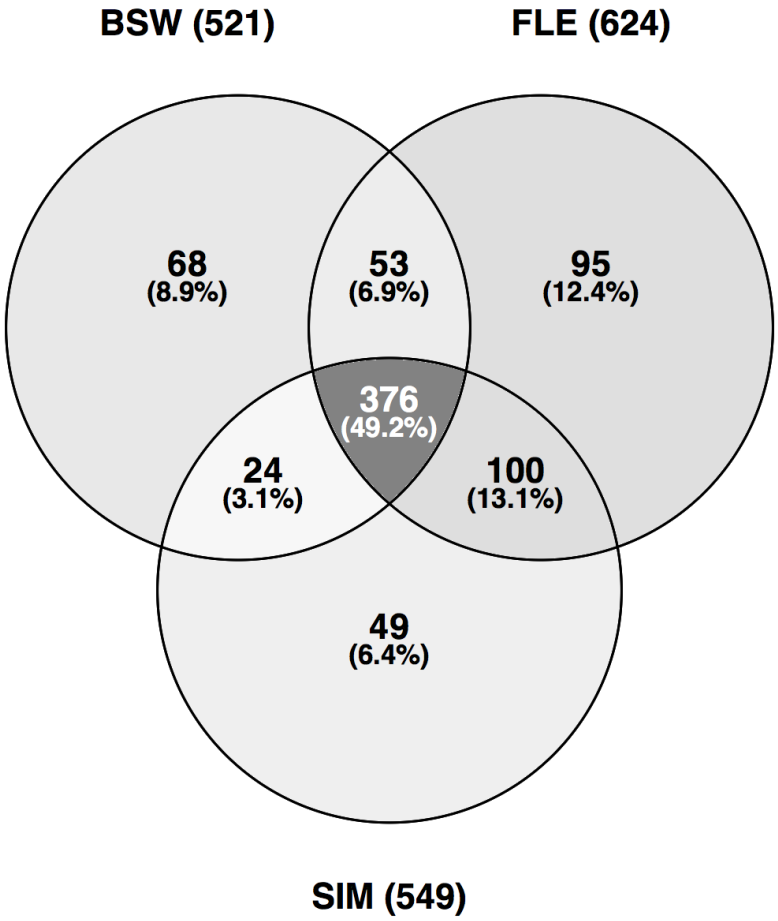

Figure 4.6 Venn diagram showing the number of High impact SNPs with minor allele frequency (MAF) >0.05 both within and across breeds, where BSW relates to Brown Swiss, FLE relates to Fleckvieh, and SIM relates to Simmental. Numbers in brackets are the total number of SNPs with MAF >0.05 in that breed.

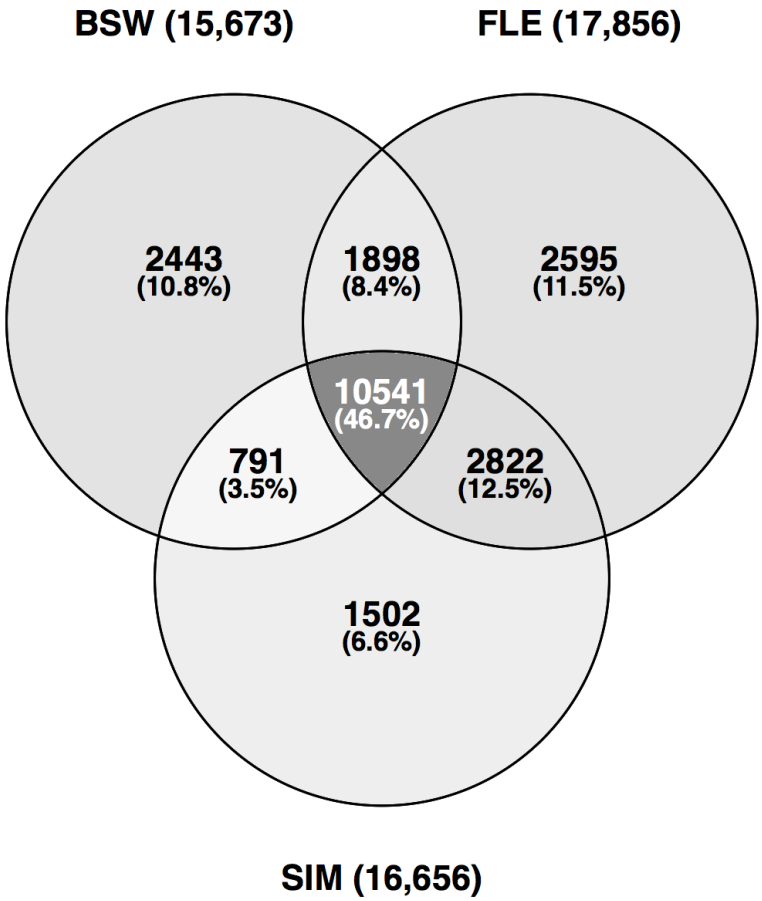

Figure 4.7 Venn diagram showing the number of Moderate impact SNPs with minor allele frequency (MAF) >0.05 both within and across breeds, where BSW relates to Brown Swiss, FLE relates to Fleckvieh, and SIM relates to Simmental. Numbers in brackets are the total number of SNPs with MAF >0.05 in that breed.

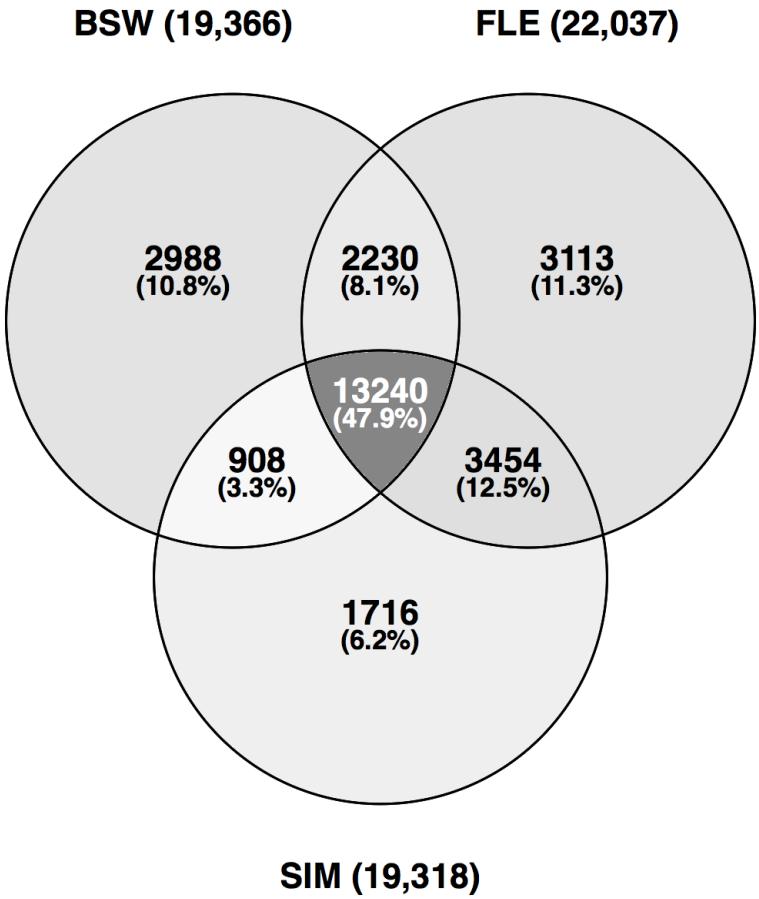

Figure 4.8 Venn diagram showing the number of Low impact SNPs with minor allele frequency (MAF) >0.05 both within and across breeds, where BSW relates to Brown Swiss, FLE relates to Fleckvieh, and SIM relates to Simmental. Numbers in brackets are the total number of SNPs with MAF >0.05 in that breed.

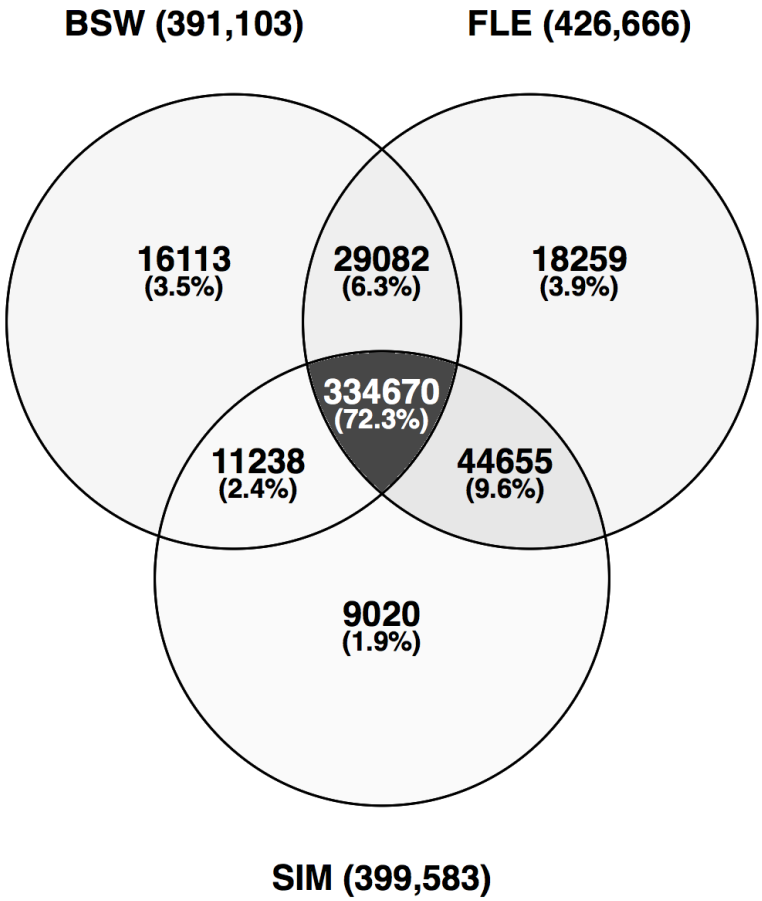

Figure 4.9 Venn diagram showing the number of BovineHD SNPs with minor allele frequency (MAF) >0.05 both within and across breeds, where BSW relates to Brown Swiss, FLE relates to Fleckvieh, and SIM relates to Simmental. Numbers in brackets are the total number of SNPs with MAF >0.05 in that breed.

4.3.2 Comparison of G Matrix calculation methods

The correlation and regression coefficients calculated when regressing the **G** matrix calculated using the VR1 method on the equivalent matrix calculated using the VR2 method are recorded in Table 4.7, with the corresponding plots in Figures 4.10 to 4.14. The regression coefficients for matrices based on filtered data suggest that for every 1% increase in relationship in a VR1 **G** matrix, there would be a corresponding estimated increase of between 1.01 and 1.03 percent if the matrix had been

calculated using the VR2 method. The non-filtered matrices showed higher regression coefficients across all categories. This increase in regression coefficient when using non-filtered data to create the **G** matrix can be attributed to the diagonal elements of **G**. Although the overall correlations using non-filtered data range from 0.98 to 1, if only the diagonal elements of the **G** matrix are considered, then the correlation between diagonals calculated using the VR1 and VR2 methods ranges between 0.79 and 0.92. The corresponding regression coefficients suggest that the relationship coefficient of an individual with itself is much lower when using the VR2 method than the VR1 method.

As the correlations between matrices using the two methods for the off-diagonal elements were close to unity, it was decided to just use one calculation method in further analyses. The VR1 method was taken forward as this method is the most common method used for research into genomic evaluations.

Table 4.7 Correlation coefficients (r) and regression coefficients (b) when regressing the relationship obtained calculating a **G** matrix using VanRaden's first method on the relationship obtained when calculating the **G** matrix based on VanRaden's second method. The **G** matrix category relates to the SNPs used to calculate the **G** matrix, where H uses High impact SNPs, HM uses High and Moderate impact SNPs, HML uses High, Moderate and Low impact SNPs, and HD uses SNPs common to the BovineHD SNP chip. Overall correlations and regressions are calculated based on all elements of **G**, whereas Diagonals only are calculated using just the diagonal elements.

G matrix	Overall		Diagonals only		Off-diagonals only	
category	r	b	r	b	r	b
H	0.98	1.06	0.79	0.47	0.98	1.26
HM	0.99	1.08	0.86	0.63	1.00	1.26
HML	0.99	1.08	0.87	0.66	1.00	1.24
HD	1.00	1.06	0.92	0.79	1.00	1.13
ALL	1.00	1.06	0.92	0.78	1.00	1.15

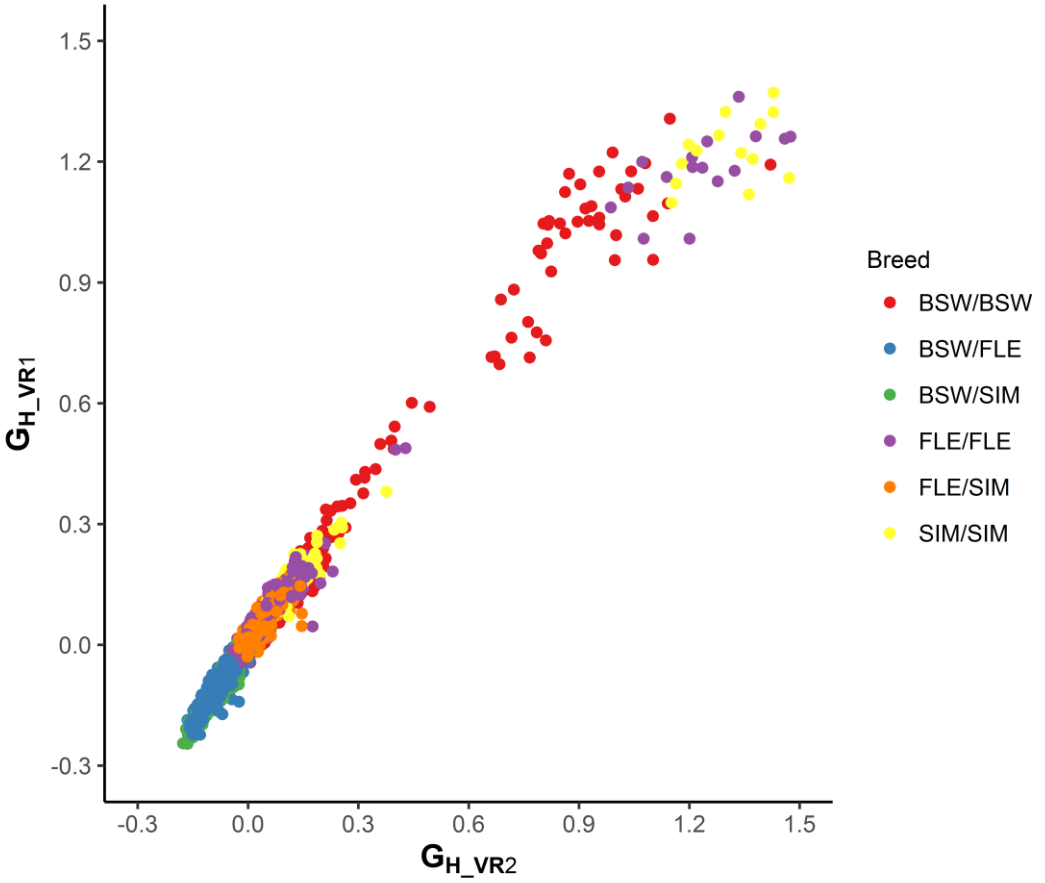

Figure 4.10 Scatterplot of relationships based on $\mathbf{G}_{\mathbf{H\_VR1}}$ plotted against $\mathbf{G}_{\mathbf{H\_VR2}}$

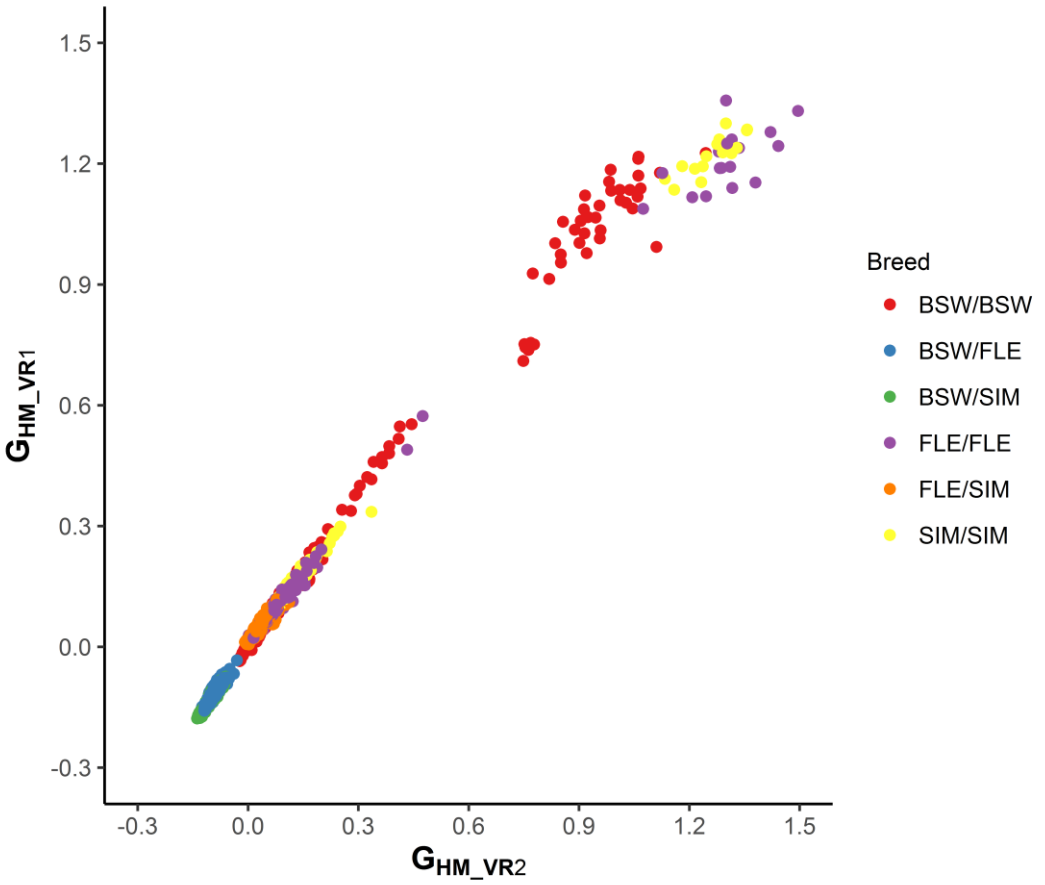

Figure 4.11 Scatterplot of relationships based on $\mathbf{G_{HM\_VR1}}$ plotted against $\mathbf{G_{HM\_VR2}}$

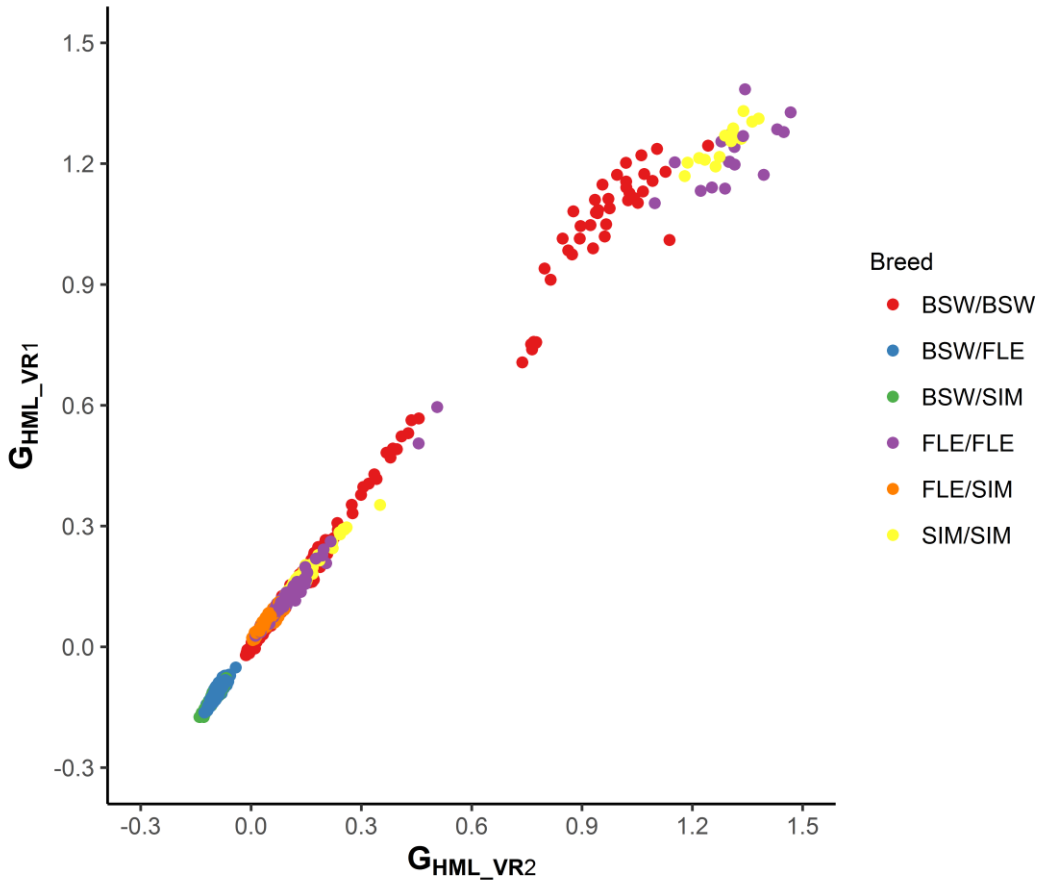

Figure 4.12 Scatterplot of relationships based on $\mathbf{G}_{\mathbf{HML\_VR1}}$ plotted against $\mathbf{G}_{\mathbf{HML\_VR2}}$

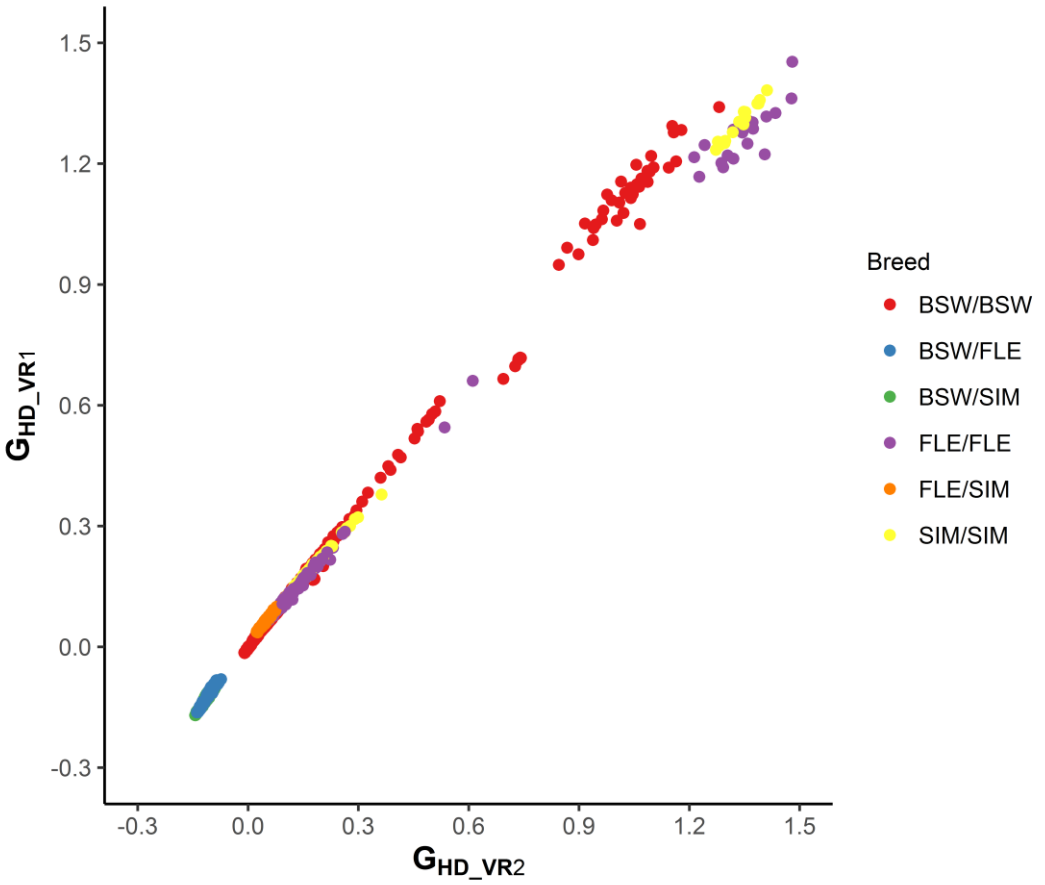

Figure 4.13 Scatterplot of relationships based on G_{HD_VR1} plotted against G_{HD_VR2}

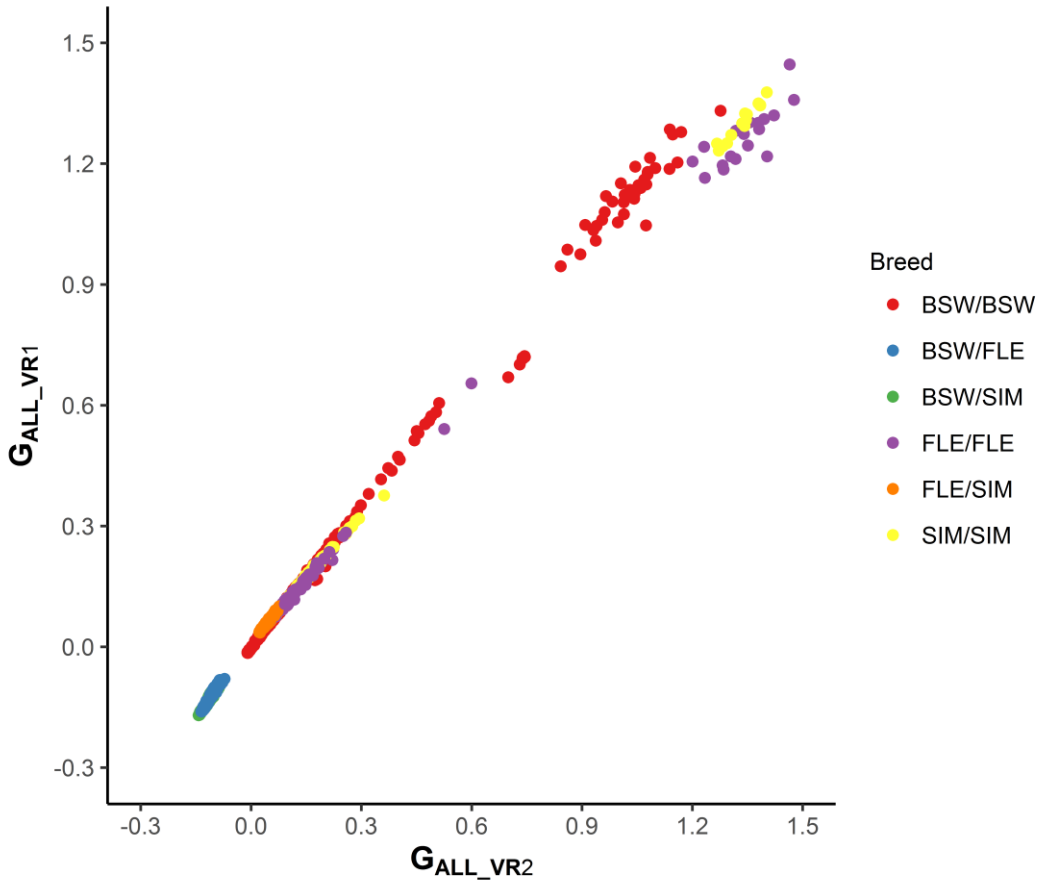

Figure 4.14 Scatterplot of relationships based on $\mathbf{G}_{ALL\_VR1}$ plotted against $\mathbf{G}_{ALL\_VR2}$

4.3.3 Principal component analyses

Plots of principal components 1 and 2 for each of the five **G** matrices are shown in Figures 4.15 to 4.19. Results of the principal components analyses were similar across matrices. In all cases the first principal component pertained to difference between breeds, and separates Brown Swiss individuals from Fleckvieh and Simmental individuals. The first principal component accounted for between 38.4% and 49.8% of the total variance between individuals, depending on the matrix analysed. The second principal component also related to differences between breed, specifically between Fleckvieh and Simmental individuals. The second principal

component accounted for between 3.7% and 4.9% of the total variance between individuals, depending on the matrix analysed. The third principal component related to further differences between breeds, and accounted for between 3.5% and 4.0% of the total variance. Increasing the number of SNPs used to calculate the **G** matrix resulted in greater separation between Fleckvieh and Simmental individuals along principal components 2 and 3.

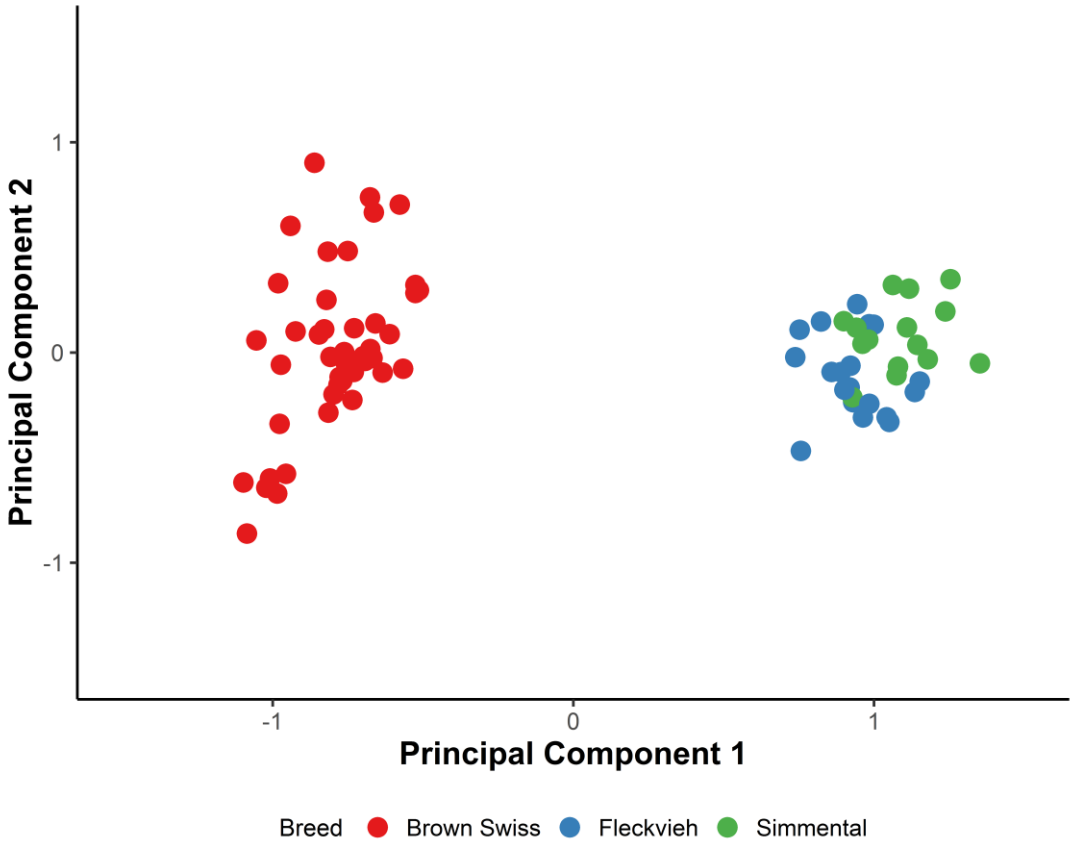

Figure 4.15 Principal components 1 and 2 based on a principal components analysis of the $\mathbf{G}_H$ matrix. Brown Swiss animals are coloured red, Fleckvieh animals are coloured blue and Simmental animals are coloured green.

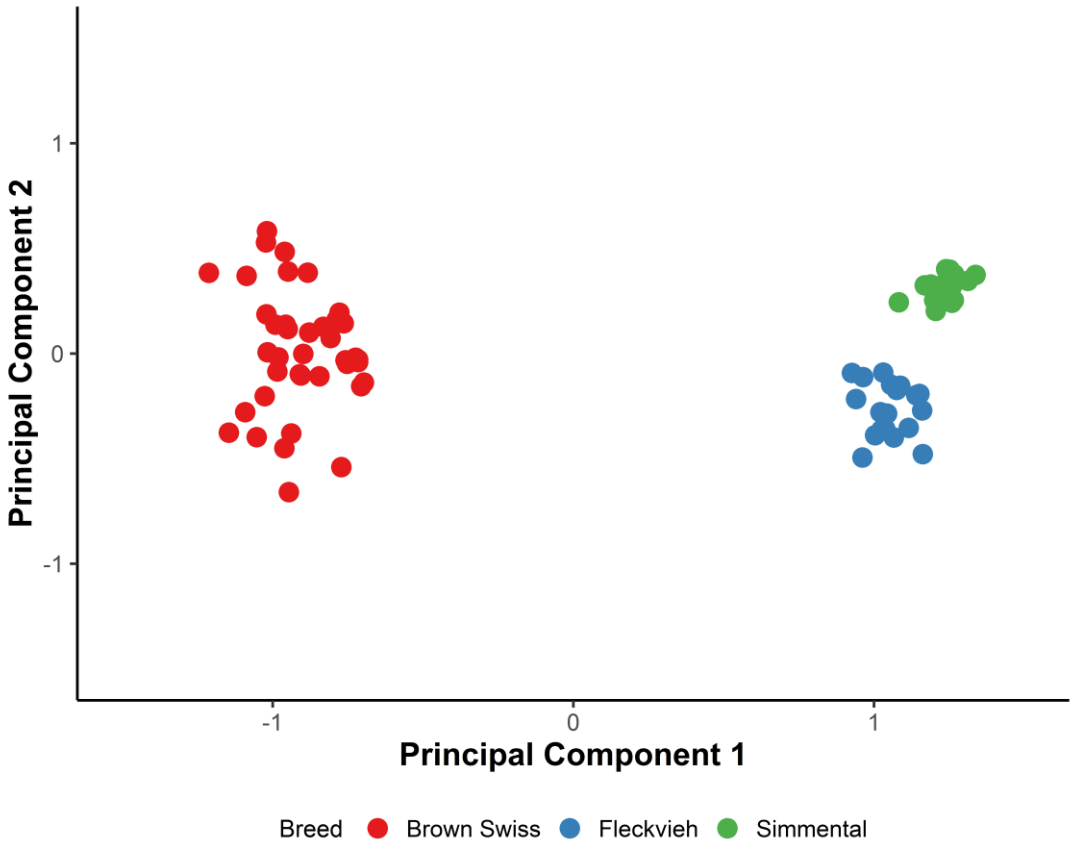

Figure 4.16 Principal components 1 and 2 based on a principal components analysis of the $\mathbf{G}_{HM}$ matrix. Brown Swiss animals are coloured red, Fleckvieh animals are coloured blue and Simmental animals are coloured green.

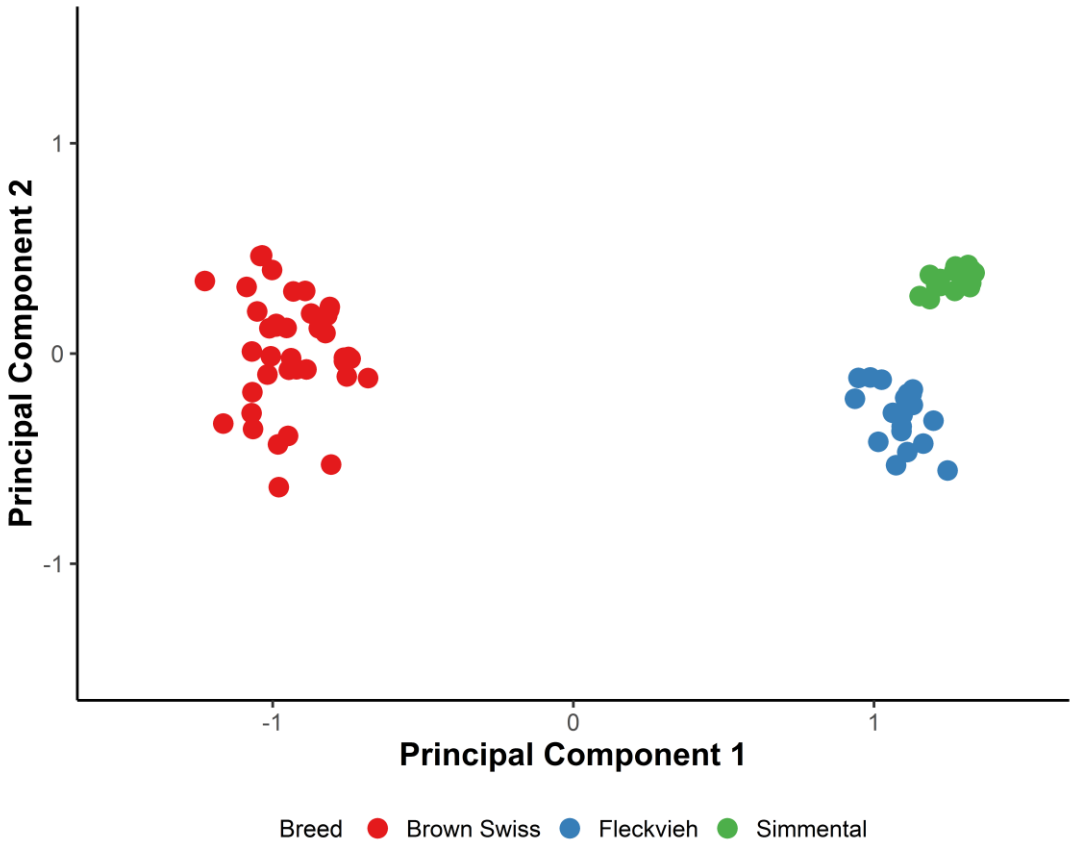

Figure 4.17 Principal components 1 and 2 based on a principal components analysis of the $\mathbf{G}_{HML}$ matrix. Brown Swiss animals are coloured red, Fleckvieh animals are coloured blue and Simmental animals are coloured green.

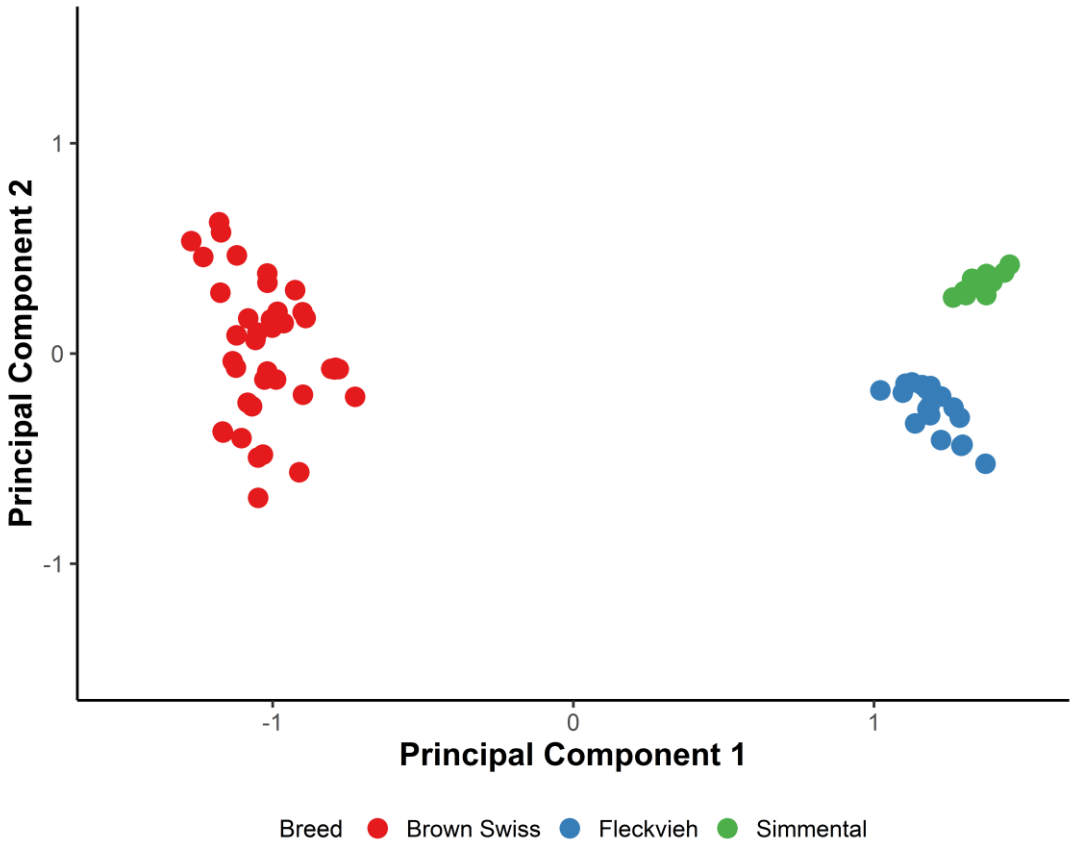

Figure 4.18 Principal components 1 and 2 based on a principal components analysis of the $\mathbf{G}_{HD}$ matrix. Brown Swiss animals are coloured red, Fleckvieh animals are coloured blue and Simmental animals are coloured green.

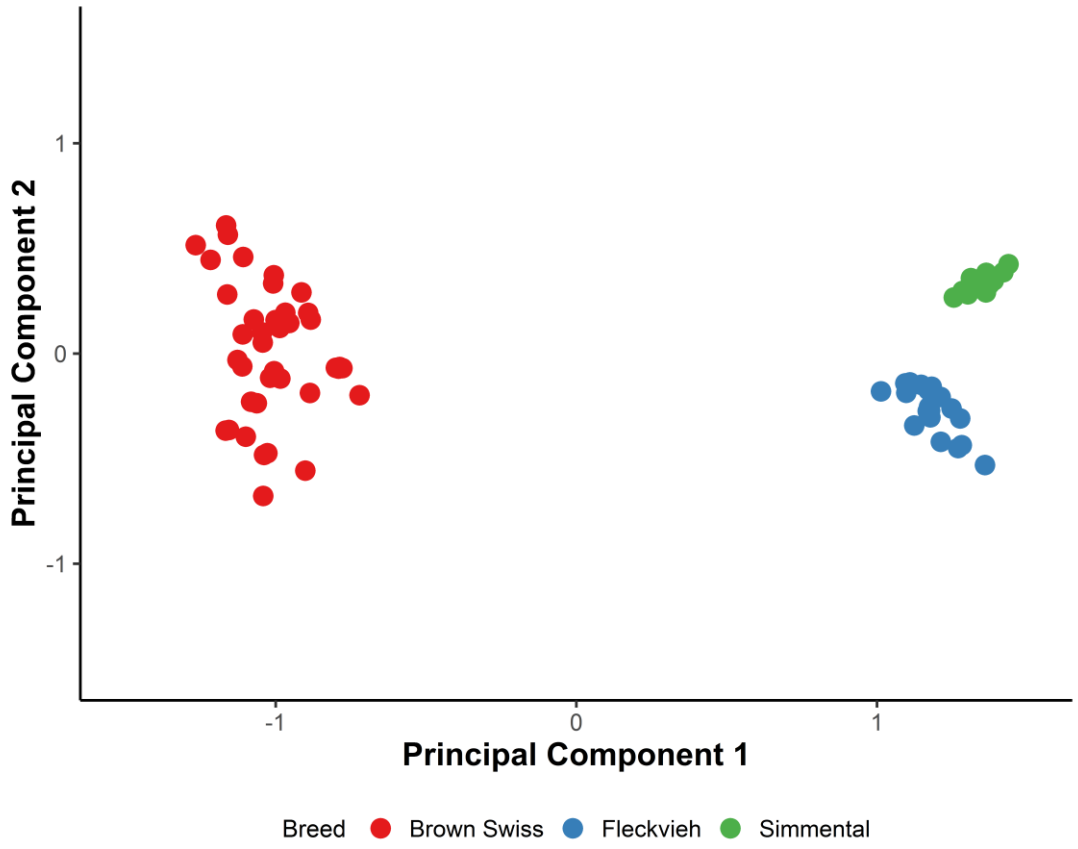

Figure 4.19 Principal components 1 and 2 based on a principal components analysis of the $\mathbf{G}_{ALL}$ matrix. Brown Swiss animals are coloured red, Fleckvieh animals are coloured blue and Simmental animals are coloured green.

4.3.4 Comparison of G matrices from different categories of SNPs

Results relating to the regression of off-diagonal relationships from $\mathbf{G}_{ALL}$ on the relationships from a) $\mathbf{G}_{H}$, b) $\mathbf{G}_{HM}$, c) $\mathbf{G}_{HML}$, and d) $\mathbf{G}_{HD}$ are displayed in Figures 4.20 to 4.23, with corresponding correlation and regression coefficients in Tables 4.7 to 4.10. For both within and between breed relationships, the correlation with $\mathbf{G}_{ALL}$ increased as more novel SNPs were used to calculate **G**. There was a significant increase in between breed correlations when including Moderate impact ($p = 0$) and Low impact SNPs ($p = 0.02$ to $p = 0$) in **G** matrix calculations. Correlations between $\mathbf{G}_{ALL}$ and **G**s from novel SNPs ranged from 0.16 to 0.82 for between-breed

relationships, depending upon the breed combination and the **G** used. The correlation for Fleckvieh-Simmental (FLE/SIM) relationships was significantly lower than for Brown Swiss-Fleckvieh (BSW/FLE) and Brown Swiss-Simmental (BSW/SIM) relationships ($p = 0.02$ to $p = 0$) when correlating $\mathbf{G}_{\mathbf{HM}}$ with $\mathbf{G}_{\mathbf{ALL}}$, and also when correlating $\mathbf{G}_{\mathbf{HM}}$ with $\mathbf{G}_{\mathbf{ALL}}$.

The results of the PCA point towards Fleckvieh and Simmental animals being more closely related to each other than to the Brown Swiss. The average between-breed relationships reflect this, being essentially zero for BSW/FLE and BSW/SIM; average relationships ranged from -0.10 to -0.12 for BSW/FLE, from -0.11 to -0.13 for BSW/SIM, and from 0.05 to 0.07 for FLE/SIM, depending on the **G** matrix.

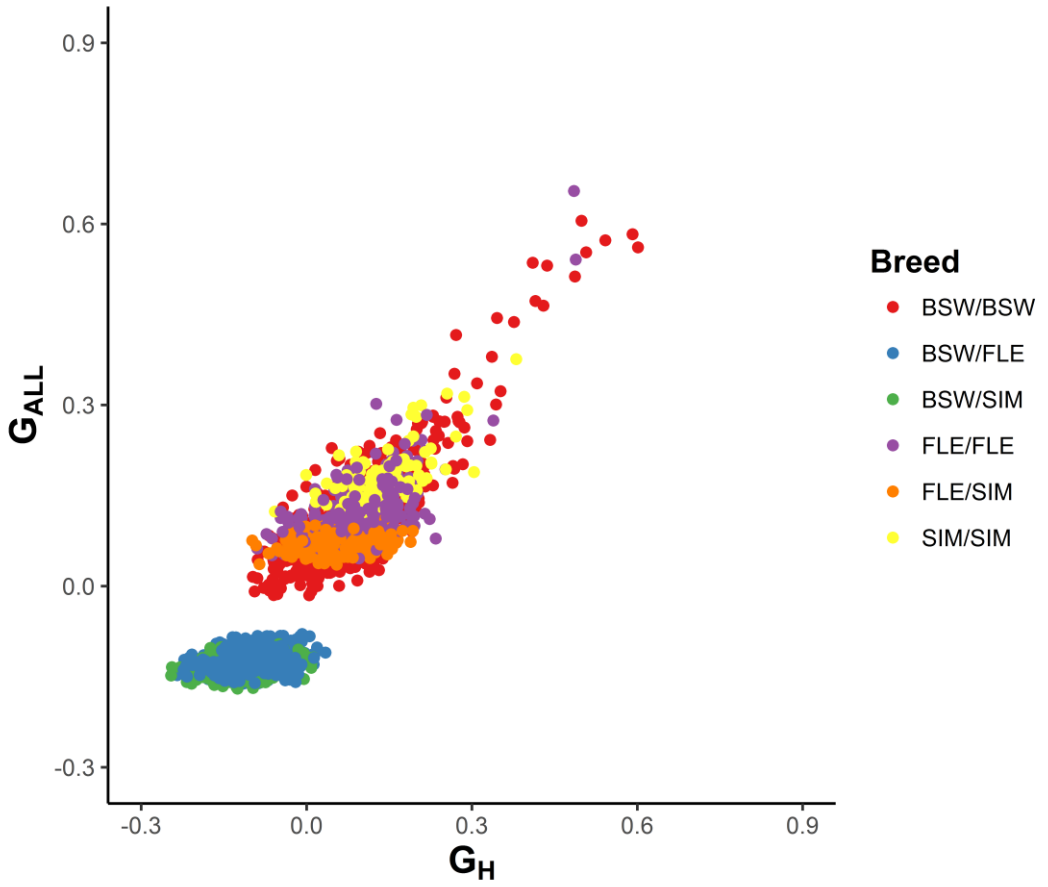

Figure 4.20 Scatter plot of relationships from $\mathbf{G}_{ALL}$ on relationships from $\mathbf{G}_{H}$

Table 4.8 Correlation and regression coefficients relating to Figure 4.15, where r is the correlation efficient and b is the regression coefficient. All refers to all relationships, BSW/BSW relates to relationships among Brown Swiss animals, BSW/FLE relates to relationships between Brown Swiss and Fleckvieh animals, BSW/SIM relates to relationships between Brown Swiss and Simmental animals, FLE/FLE relates to relationships among Fleckvieh animals, FLE/SIM relates to relationships between Fleckvieh and Simmental animals, and SIM/SIM relates to relationships among Simmental animals.

Relationship	r	b
BSW/BSW	0.85	0.79
BSW/FLE	0.25	0.09
BSW/SIM	0.16	0.06
FLE/FLE	0.64	0.51
FLE/SIM	0.24	0.06
SIM/SIM	0.63	0.43

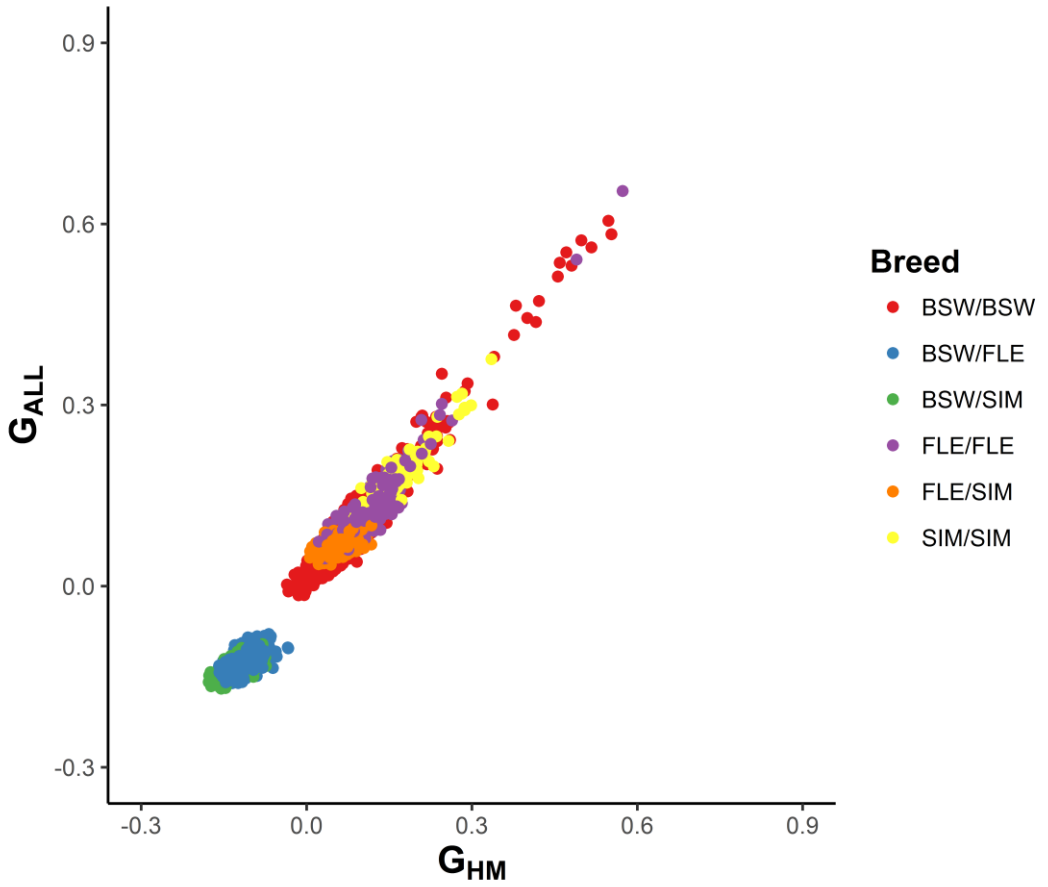

Figure 4.21 Scatter plot of relationships from $\mathbf{G_{ALL}}$ on relationships from $\mathbf{G_{HM}}$

Table 4.9 Correlation and regression coefficients relating to Figure 4.16, where r is the correlation efficient and b is the regression coefficient. All refers to all relationships, BSW/BSW relates to relationships among Brown Swiss animals, BSW/FLE relates to relationships between Brown Swiss and Fleckvieh animals, BSW/SIM relates to relationships between Brown Swiss and Simmental animals, FLE/FLE relates to relationships among Fleckvieh animals, FLE/SIM relates to relationships between Fleckvieh and Simmental animals, and SIM/SIM relates to relationships among Simmental animals.

Relationship	r	b
BSW/BSW	0.97	1.04
BSW/FLE	0.70	0.59
BSW/SIM	0.74	0.56
FLE/FLE	0.93	0.99
FLE/SIM	0.61	0.35
SIM/SIM	0.92	0.88

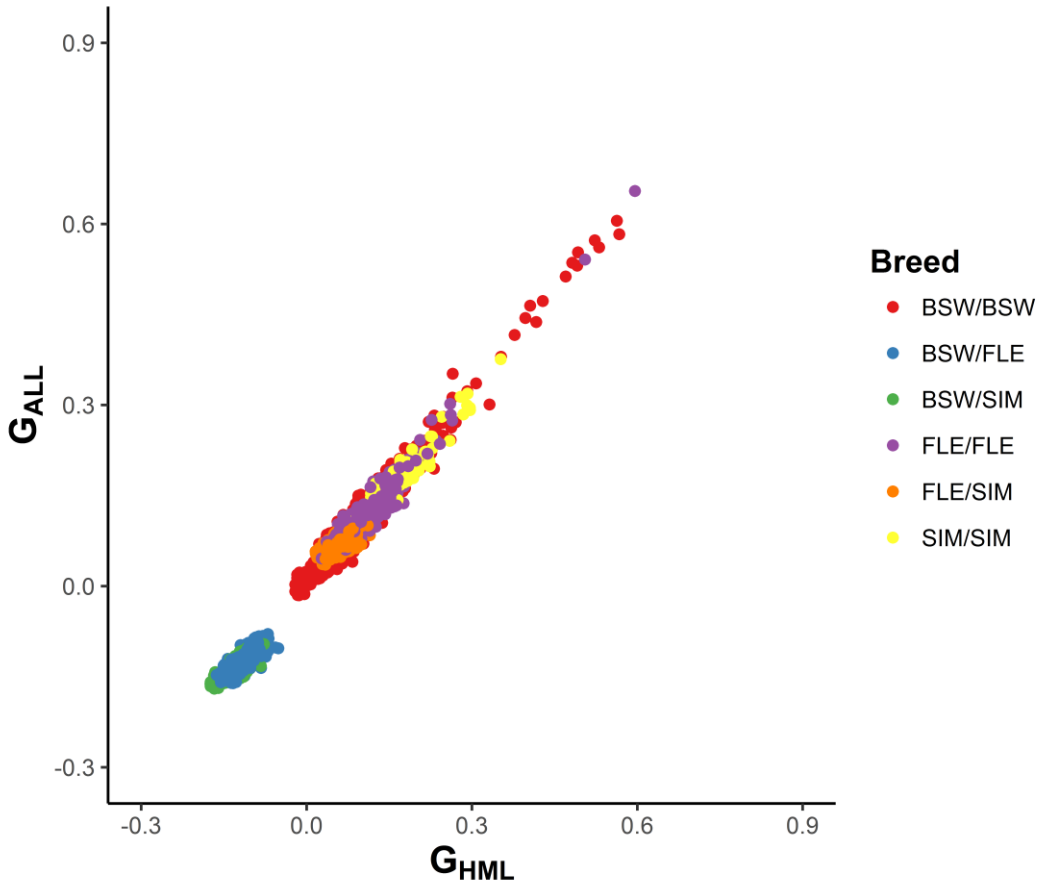

Figure 4.22 Scatter plot of relationships from $\mathbf{G_{ALL}}$ on relationships from $\mathbf{G_{HML}}$

Table 4.10 Correlation and regression coefficients relating to Figure 4.17, where r is the correlation efficient and b is the regression coefficient. All refers to all relationships, BSW/BSW relates to relationships among Brown Swiss animals, BSW/FLE relates to relationships between Brown Swiss and Fleckvieh animals, BSW/SIM relates to relationships between Brown Swiss and Simmental animals, FLE/FLE relates to relationships among Fleckvieh animals, FLE/SIM relates to relationships between Fleckvieh and Simmental animals, and SIM/SIM relates to relationships among Simmental animals.

Relationship	r	b
BSW/BSW	0.98	1.04
BSW/FLE	0.81	0.75
BSW/SIM	0.82	0.69
FLE/FLE	0.96	1.00
FLE/SIM	0.70	0.49
SIM/SIM	0.95	0.91

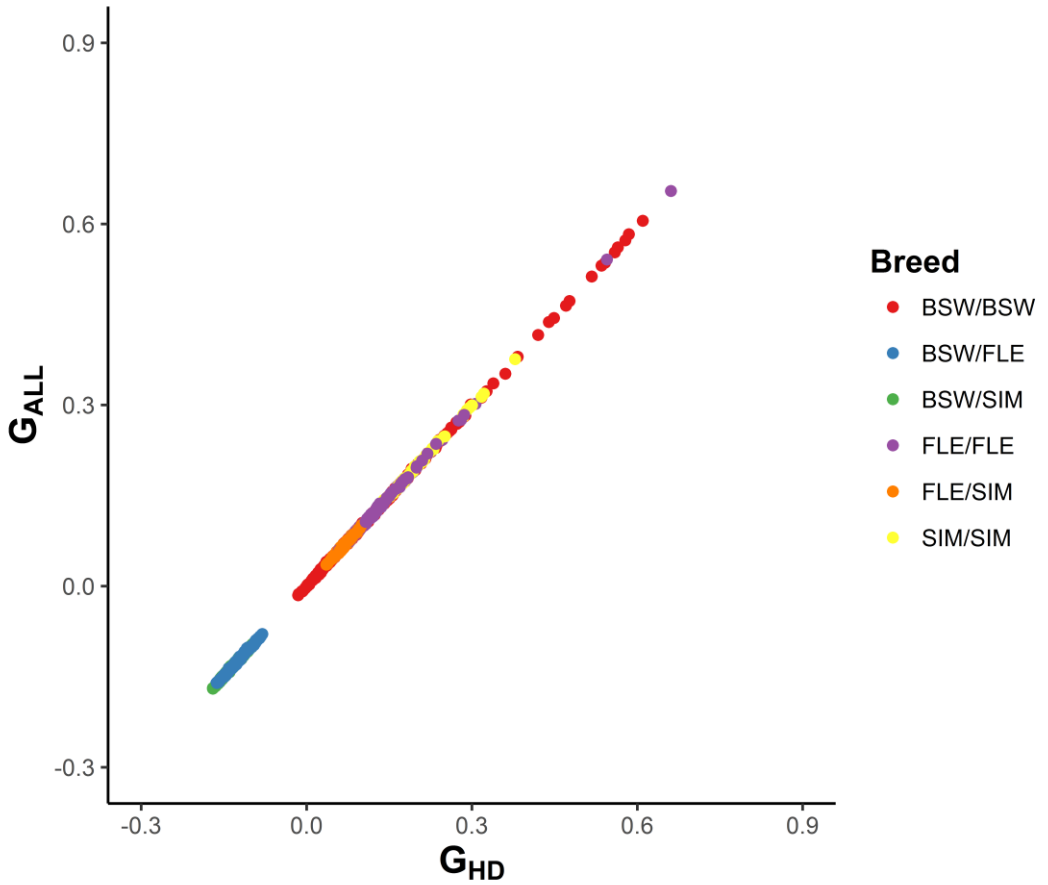

Figure 4.23 Scatter plot of relationships from $\mathbf{G_{ALL}}$ on relationships from $\mathbf{G_{HD}}$

Table 4.11 Correlation and regression coefficients relating to Figure 4.18, where r is the correlation efficient and b is the regression coefficient. All refers to all relationships, BSW/BSW relates to relationships among Brown Swiss animals, BSW/FLE relates to relationships between Brown Swiss and Fleckvieh animals, BSW/SIM relates to relationships between Brown Swiss and Simmental animals, FLE/FLE relates to relationships among Fleckvieh animals, FLE/SIM relates to relationships between Fleckvieh and Simmental animals, and SIM/SIM relates to relationships among Simmental animals.

Relationship	r	b
BSW/BSW	1.00	0.99
BSW/FLE	1.00	0.98
BSW/SIM	1.00	0.99
FLE/FLE	1.00	0.99
FLE/SIM	0.99	0.99
SIM/SIM	1.00	1.00

4.4 Discussion

4.4.1 Allele sharing across breeds

The proportion of SNPs with low-MAF (<0.05) among High, Moderate and Low impact SNPs was approximately threefold higher in comparison to the proportion of low-MAF SNPs common to the HD chip. These proportions are similar to those observed by Eynard et al. (2015) when comparing SNPs from sequence to SNPs from genotyping chips. We would expect the proportion of low-MAF SNPs to be higher for novel SNPs, as the SNPs on genotyping chips such as the Illumina BovineHD are specifically chosen because they have been proven to be highly polymorphic across a large number of breeds (Matukumalli et al., 2009), whereas one may expect that SNPs with a significant impact on protein structure, such as non-synonymous mutations in coding regions, are likely to be present at lower frequencies in the genome, as functional elements of the genome are generally more highly conserved. A higher number of SNPs with MAF >0.05 was observed for Fleckvieh individuals than for Brown Swiss and Simmental, which reflects the higher variability seen in the Fleckvieh population by Sánchez-Molano et al (2016). This is illustrated well in Figure 4.3, where we observe a difference between allele frequency in Brown Swiss and Fleckvieh breeds for High, High Moderate, and High Moderate Low SNPs, where the reference allele frequency is fixed at 1 in Brown Swiss, but is lower than 1 in the Fleckvieh breed.

The proportion of alleles segregating across two breeds (Figures 4.3 to 4.5) did not necessarily equate to higher relationship coefficients between breeds, with a lower average relationship coefficient observed for BSW/FLE and BSW/SIM relationships

using $\mathbf{G_{ALL}}$ than $\mathbf{G_H}$, despite a higher percentage of alleles from ALL SNPs segregating across the two breeds. When mean reference allele frequencies were compared between breeds, a higher proportion of SNP bins were closer to fixation or fixed in the Brown Swiss breed than either the Fleckvieh or the Simmental (Figures 4.3 to 4.5). Lower differences were observed between mean reference allele frequency for Fleckvieh and Simmental. This decrease in average relationships when ALL SNPs are used to calculate **G** could therefore be due to SNPs having a higher minor allele frequency in Fleckvieh or Simmental than we see in Brown Swiss.

4.4.2 Comparison of G-matrix calculation methods

Few published studies have directly compared the resulting relationship matrices calculated using VR1 and VR2, with those that have demonstrating little difference between the two (Lorenz et al., 2015). The main difference between the two methods is how they incorporate information on allele frequencies, with VR1 being scaling by a single value, $\sum 2p_i(1-p_i)$, and VR2 normalising the SNPs by dividing by $2p_i(1-p_i)$ locus by locus. The result of this difference is that the effect of low-MAF loci is shrunk more using VR1 than using VR2, and so low-MAF loci have a greater influence on the relationship when calculated using VR2 than VR1. It therefore stands to reason that we see a difference between the two relationship matrices, as rare alleles have been included in our data set rather than being filtered out. Interestingly, the differences between the two matrices were mainly due to the diagonal elements, with VR2 underestimating the diagonals compared to VR1. Both the VR1 and VR2 matrices are scaled to have a mean of zero, hence the results in Table 4.7, where we observe regression coefficients below 1 for the diagonal

elements of the matrix, and above 1 for the off-diagonal elements, but these elements do not affect subsequent evaluations (Tier et al., 2015). As the VanRaden study (2008) was based on simulated data, they were able to compare the diagonal elements of G (the so-called "genomic inbreeding coefficient") to the true values, and showed that the estimates based on the VR1 method were more precise than those using the VR2 method, which contributed to our decision to use only the VR1 method in further analyses.

4.4.3 Between breed correlations

The focus of this study is whether SNPs extracted from WGS data could be useful in estimating genomic relationship between individuals of different breeds, and whether using SNPs with a "significant" putative impact of the genome could improve the accuracy of multi-breed genomic evaluations. As such, the remainder of this discussion will concentrate on the correlations of between-breed relationships.

As expected, the correlations of between-breed relationships between the different **G** matrices were lower than that seen within breeds. Correlations between $\mathbf{G_{HM}}$ and $\mathbf{G_{ALL}}$ were significantly higher than correlations between $\mathbf{G_{H}}$ and $\mathbf{G_{ALL}}$ for all three breed combinations ($p = 0$), and significant increased again when Low impact SNPs were included in calculating G ($\mathbf{G_{HML}}$). These correlations are in agreement with a study by Rolf et al (2010), who investigated the use of reduced marker panels in calculating a **G** matrix in a population of Angus cattle. The results of the Rolf et al. study suggested that marker panels of approximately 10,000 SNPs were sufficient to accurately estimate genomic relationships between individuals, however using a

marker panel made up of less than 2,500 SNPs are not appropriate for estimation of G matrices as they are too sensitive to changes in sample size.

For all regressions of $\mathbf{G_{ALL}}$ on **G**s from novel SNPs, the slope of the regression was well below 1, therefore considering our assumption that $\mathbf{G_{ALL}}$ represents the most accurate estimate of relationships between individuals, relationships between breeds are generally overestimated when calculating **G** from novel SNPs obtained from WGS data.

The correlations for FLE/SIM relationships between **G**s created using different sets of novel SNPs and $\mathbf{G_{ALL}}$ were significantly lower than for BSW/FLE and BSW/SIM relationships. This was unexpected considering that both the PCA and analysis of allele frequencies demonstrate that Fleckvieh and Simmental are more closely related to each other, than to Brown Swiss. The most likely reason for this is that although the difference between FLE/SIM and other between breed correlations is statistically significant, the lower correlations could simply be a product of chance, due to small sample size. In a study looking at ancestral haplotypes in Brown Swiss, Fleckvieh and Simmental, Sanchéz-Molano et al. (2016) suggested that the two breeds are not as similar as expected, and that the flow of genetic material between the two breeds is uneven, with a higher flow of genetics from Simmental to Fleckvieh than the other way around. In the present study, all Simmental individuals originate from Switzerland, whereas the Fleckvieh individuals originate from Austria and Germany, with none sampled from Switzerland. It may be possible that the influence of the local Bavarian breeds is higher in Austrian and German Fleckvieh, and so our two

sample populations are not as similar as we may first expect. Another possibility could be that the relationships between Fleckvieh and Simmental are not captured as effectively using the sets of novel SNP markers as they are using $\mathbf{G}_{\mathbf{ALL}}$.

4.4.4 Randomly selected vs "significant" SNPs

The main limitation of this study has been the small sample size, which has restricted us to comparing the genomic relationship matrices using different subsets. Genotype and phenotype data was available for approximately 600 animals across the three breeds, and the initial aim of the study was to use the sequenced animals as a reference population to allow imputation of novel SNPs into the genotyped population, and subsequently carry out a multi-breed genomic evaluation based on the different SNP subsets. However, for accurate imputation to be possible, there must be enough parent-offspring pairs in the reference population to allow correct haplotype phasing. Only five parent-offspring relationships were present between our sequenced animals, and therefore we were unable to implement the imputation.

A recent study by Van den Berg et al. (2016) simulated causal variants and phenotypes to investigate the utility of sequence data in across-breed genomic evaluations as opposed to 50k or HD chip data using the GBLUP method. The results showed that using sequence variants that are close to (and therefore in high LD with) causal variants for a trait improves the accuracy of prediction in comparison to both the 50k and the HD SNP chips, and also that using only those SNPs closest to the causal variants on each of the SNP chips resulted in higher accuracy of GEBVs than using all SNPs.

The hypothesis behind selecting SNPs based on their putative impact with regards to protein structure and behaviour, is that SNPs in functional regions of the genome are more likely to be close to the causal variants, or potentially be causal mutations themselves. If this is the case, based on the results of our analysis and the results presented by Van den Berg et al., we would suggest that using panels of "significant" SNPs for genomic evaluations should improve the accuracy of prediction in a multi-breed scenario, but genomic evaluations should be carried out on a larger data set to fully test this hypothesis.

4.4.5 Conclusion

We have shown that SNP variant predictors are capable of extracting SNPs that are polymorphic across breeds, and that we can use them to predict relationships that correlate well with relationships estimated using a higher density of markers.

To test the hypothesis that using SNPs with a "significant" putative impact on the genome will improve the accuracy of multi-breed genomic evaluations, a larger data set will need to be identified to allow the estimation of GEBVs from different SNP subsets. Analysis of multiple traits would be of interest, as would comparing different methods of prediction, e.g. GBLUP and BayesC, if the traits analysed have differing genetic architecture. Examination of SNP effects from a Bayesian analysis could give some insight as to whether the novel SNPs have a higher effect on the trait of interest than SNPs from current genotyping chips.

Chapter 5: General Discussion

5.1 Introduction

The introduction of genomic prediction into the dairy cattle industry has caused an industry-wide revolution in dairy cattle breeding (Taylor et al., 2016). Particularly within the Holstein breed, this technology has allowed breeders to increase the rate of genetic gain per annum by 50% in the past 7 years, reducing the generation interval by marketing young bulls with genomic EBVs for breeding, and allowing differentiation between full-sibs (García-Ruiz et al., 2016).

Since the introduction of genomic selection to the present day, its application is still limited to those breeds for which a large reference population of animals with both phenotypes and genotypes is available. Across-breed genomic evaluations, both multi-breed and crossbred, have yet to be widely implemented at a commercial level.

This thesis has sought to explore the potential to implement both multi-breed and crossbred genomic predictions using methods of prediction currently implemented commercially in the UK.

5.2 Thesis Overview

In attempting to understand the potential for across-breed genomic selection to be implemented in dairy cattle, this study first concentrated on the multi-breed evaluation scenario, in which we sought to improve the accuracy of genomic evaluations in the numerically small British Friesian breed, by augmenting the size of the reference population using genotypes and phenotypes from Holstein bulls. Both production and non-production traits were considered, and the results showed

that in 15 of the 20 analyses across the five traits, incorporating Holstein genotypes into the reference population improved the accuracy of genomic prediction in Friesians by between 1% and 39%. More promisingly, we showed that implementing the single-step method, which allows the incorporation of further phenotypes into the reference population, the accuracy of evaluation significantly increased for Friesian animals. We also tested the utility of a HD marker panel in a multi-breed genomic prediction, but saw no consistent improvement in the accuracy of evaluation when more markers were used to estimate genomic relationships.

Chapter 3 considered the second interpretation of an across-breed genomic evaluation, and an African crossbred reference population of animals was used to estimate GEBVs in crossbred selection candidates. Analyses were limited to one trait – milk yield – due to low levels of data recording in Sub-Saharan Africa, but the accuracies obtained suggested that implementation of genomic selection in a highly cross-bred reference population may be possible if targeted recording and genotyping measures were to be taken.

Chapter 4 again considered the multi-breed evaluation scenario, but focused on the use of sequence data for the estimation of genomic relationships across breeds, specifically whether it is possible to accurately estimate genomic relationships between individuals based on SNP variants that have a "significant" putative impact on the genome with regards to how they affect biological functions such as protein structure or behaviour. Genomic relationship matrices estimated using all "significant" SNPs were highly correlated with genomic relationship matrices based

on all available SNPs, suggesting that it may be possible to use these SNPs for genomic evaluation; however this proposed method has yet to be substantiated via GEBV estimation.

This chapter will discuss results from the previous three chapters, along with the current situation regarding implementation of multi-breed and crossbred genomic evaluations in the UK dairy industry, along with some thoughts on further work.

5.3 Application of multi-breed genomic evaluations in the UK

The results of chapter 2 show that incorporating more animals into a multi-breed reference population does improve the accuracy of a genomic prediction, which is in line with results from other studies (Erbe et al., 2012; Weber et al., 2012; Hozé et al., 2014; Zhou et al., 2014b). Although the accuracy observed for Friesian GEBVs is, as expected, lower than traditional EBVs for proven bulls with many daughters, it is higher than the parent average value. This improvement in accuracy is exactly why genomic selection has been successful, because young bulls can be selected and marketed at sexual maturity instead of having to go through progeny testing before selection as an elite sire (Lillehammer et al., 2011). The difference in accuracy between a genomic young bull and a proven bull is approximately 20% (García-Ruiz et al., 2016), but the increase in the rate of genetic gain per annum due to the reduced generation interval when using young bulls makes up for this difference. As the Friesian GEBVs calculated in this study have a higher accuracy than parent average, we can conclude that it would be of interest to use a multi-breed reference population to implement genomic evaluations for the British Friesian breed. The British Friesian

is a numerically small breed, with only approximately 12,500 cows currently in the UK national herd (F Pearston, AHDB Dairy, personal communication), and so it is critical that any selection that takes place within the Friesian breed is carefully thought out to maintain levels of genetic diversity and limit inbreeding. As genomic selection is based on genomic relationship matrices, it captures Mendelian sampling variance, enabling breeders to distinguish between full-sibs when making breeding decisions. Furthermore, simulation studies suggest that using genomic relationships as opposed to pedigree based relationships result in lower increases of co-ancestry over generations (Rodríguez-Ramilo et al., 2015; Bastiaansen et al., 2012). The results of these studies suggest that implementation of genomic selection could be more suitable for Friesian breeders than traditional pedigree-based selection. Control of inbreeding would still be necessary however, and so the technique should be combined with methods such as optimum contributions selection to facilitate this (Sonesson et al., 2012).

We made every effort to be exacting when choosing a data set to carry out this study, as we wanted the data set to be as clean as possible in order to effectively test our hypothesis and avoid spurious results. Our results in Friesians are based on an extremely small validation population, and so the margin for error was large. We would therefore be cautious about promoting the publication of Friesian GEBVs based purely on these results. When it comes to commercial evaluations, all available data tends to be included in the analysis, and so genotypes would be available for some animals that were discarded from this data set. After the work in chapter 2 was completed, GEBVs for milk yield were estimated for 46 Friesian bulls using a

reference population containing all Holsteins used for national genomic evaluations, along with some Friesian genotypes. These GEBVs were calculated using SNPs from the 50k chip. The accuracy of Friesian GEBVs in this scenario was 0.70, with a regression slope of 0.92. These results demonstrate that when using a larger reference population that primarily represents the Holstein breed, we see further improvement in prediction accuracy for milk yield compared to the accuracies achieved in chapter 2. The results based on national data were obtained using the SNP-BLUP method, which is equivalent to the GBLUP method. We believe that further gains could be achieved were additional phenotypes to be incorporated into the evaluation via the HBLUP method. The HBLUP method is not currently implemented for dairy cattle evaluation in the UK, but the same team has used the technique for calculating GEBVs in UK dairy goats (Mucha et al., 2015), and so it should be straightforward to implement the same software and procedures in dairy cattle genomic evaluations as is currently used in the UK dairy goat population.

We saw higher accuracies of evaluation in our Holstein animals using the HD chip than has been seen using HD genotypes in previous genomic selection studies (Su et al., 2012; Erbe et al., 2012), which may have been due to the accuracy of imputation, either due to the size of the reference population procedure, or the software used for imputation. Previous studies looking at the size of reference population with regards to imputation accuracy (Ma et al., 2014; Druet et al., 2010), have showed that the accuracy of imputation does increase as the reference population size increased. Druet et al. (2010) saw a 1% increase in accuracy when moving from using 1,000 individuals to 2,000 individuals in a Dutch Holstein reference population, and Ma et

al. (2014) saw a 3% increase when moving from a reference population of approximately 1,750 individuals to 4,398 individuals, where the smaller reference population contained Chinese Holsteins and the larger contained both Chinese and Nordic Holsteins. Studies comparing imputation accuracy based on different software (Jattawa et al., 2016; Brøndum et al., 2014; Weng et al., 2013), showed that the accuracy of imputation when using findhap was consistently lower than using FImpute, and in two out of three studies findhap was also outperformed by BEAGLE software. Imputation is currently carried out for UK genomic evaluations using findhap due to the run time being faster than other software. However, the results in chapter 2 suggest that it may be worth carrying out further testing of findhap compared to other imputation software for this data set.

All this is positive news with regards to the implementation of Friesian genomic evaluations, but the UK also hopes to roll out evaluations for other numerically small breeds such as the Ayrshire and the Guernsey. Thanks to recent data sharing agreements with countries such as the USA and Canada, the UK has access to larger numbers of genotypes from Ayrshire (n = 2,239) and Guernsey (n = 2,489). Of the Ayrshires, 88.2% have been genotyped using chips containing 50k or more markers, as opposed to 33% of Guernseys. Unfortunately, the data sharing agreements that allowed access to these genotypes happened too recently to allow the feasibility of multi-breed Guernsey or Ayrshire genomic evaluations to be addressed within this thesis. However, owing to the larger number of genotypes available for these breeds, we would be hopeful that using multi-breed reference populations to inform evaluations for these breeds is feasible. As a high proportion of the Guernsey

genotypes available are genotyped at low density (10k – 20k SNPs), imputation methods would need to be implemented for the Guernsey population, and as discussed above, care would need to be taken with regards to the method used to ensure high imputation accuracy. Whether we would also see improvements in accuracy achieved for these breeds however, also very much depends on factors such as the level of LD between markers and QTL, whether the marker and QTL are in the same phase across breeds, and whether there are differences in allele frequencies across the breeds of interest (De Roos et al., 2009; Kizilkaya et al., 2010; Goddard et al., 2015). These questions remain to be addressed in future studies. A multi-breed reference population for genomic evaluation of numerically small breeds in the UK is likely to be predominantly made up of Holsteins, but the selection candidates will be the smaller breed, e.g. the Ayrshire. If the allele frequency of a QTL in Ayrshire is much higher than the allele frequency in Holstein, for example, the effect of the QTL will not be estimated well, despite having a high effect on the genetic variance of the trait in Ayrshires (Goddard et al., 2015). It has been proved that some QTL of large effect are present at different frequencies across breeds. Examples of this occurring include the DGAT1 mutation (which has an effect on fat content in milk) being observed at varying frequencies in different Italian cattle breeds, and was missing in two of them (Scotti et al., 2010). The impact of these differences in allele frequency on the accuracy of multi-breed genomic evaluations depends on the size of the difference in allele frequency, and the proportion of genetic variance explained by the QTL.

5.4 Application of crossbred genetic evaluations in the UK

The results of chapter 3 present a good opportunity to implement selection within crossbred African cattle populations, as the lack of reliable pedigree information for these animals restricts the use of traditional genetic evaluation methods in the African smallholder system. The work carried out was based on a small data set, with phenotypic data only available for milk yield. The cows analysed in this study had differing proportions of exotic (imported) dairy genetics, and this was reflected in their yields, with those cows with higher proportions of exotic dairy genetics producing more milk. If genomic evaluation is to be properly implemented in this African population, care will need to be taken to ensure that we do not inadvertently select for cows with higher proportions of imported dairy genetics due to them having a higher genetic merit for yield. The dairy cattle industry in developed countries such as the UK has suffered in the past due to the use of narrow breeding goals, which have led to a decrease in fitness traits such as fertility and lifespan. Lessons have been learned, and breeding goals have been widened in order to try and breed better "all-round" dairy cows, by developing a selection index known as the "profitable lifetime index" (PLI) that gives more weight to health and fitness traits than previously. Before genomic evaluations can be fully implemented in the African dairy smallholder system, strong foundations must be laid by implementing appropriate targeted recording techniques, to ensure that sufficient high quality phenotypic data can be collected not only for production traits, but also for fitness traits in order to assemble a selection index that is suited to the system that the cows will ultimately be managed in.

At first glance, the fact that we saw positive results in our African population gives hope that crossbred genomic evaluations could be implemented for dairy cattle in the UK. However, a previous study in beef cattle where there was access to pedigree data (Mujibi et al., 2011) did not see such promising results, mainly due to the fact that the accuracy of a genomic evaluation did not sufficiently exceed that of a pedigree-based genetic evaluation. The other factor that needs taking into account is the genetic make up of the crossbred population of interest. The population analysed in chapter 3 was highly crossbred, and though the PCA showed that the animals could be grouped according the proportion of exotic dairy genetics, the population overall looked to be homogeneous, with no separation between groups. This suggests that so much crossbreeding has been carried out within this population, that it resembles a composite breed.

The primary focus on crossbred genomic evaluations involving the UK dairy industry would be to improve carcass related traits in beef crossbred offspring of dairy animals, as selling crossbred calves into the beef supply chain (either to be finished for beef production or for use as a suckler cow) increases returns for dairy farmers (Vickers et al., 2014). Due to the nature of the UK dairy industry where the majority of animals are purebred, crossbred offspring of dairy animals would most likely be F1 crosses (i.e. the offspring of two purebred animals of different breeds). Estimating genetic merit of an F1 individual compared to a highly crossbred individual is likely to be more difficult, as heterosis will have more of an effect in the F1 crossbred animal. Heterosis and recombination loss are currently accounted for in UK dairy across breed genetic evaluations by making use of breed proportion

information calculated using the entire national pedigree. Animals are grouped into 4 classes – Holstein, Friesian, Channel Island and Others – and heterosis and recombination loss is calculated for all combinations of these classes.

Crossbred genomic evaluations in the UK would be likely to concentrate on predicting crossbred performance GEBVs for purebred dairy cattle, i.e. A GEBV that will give an estimate as to the performance of a cow's crossbred offspring. A number of studies in pigs (Esfandyari et al., 2015; Hidalgo et al., 2015) suggest that a crossbred reference population should be used to predict GEBVs for crossbred performance in purebred animals, as they result in higher accuracies of prediction. This leads us to a further problem. Due to associated costs, commercial genotyping is generally only carried out for animals of high value in a population, and in the case of the UK dairy industry these animals tend to be purebred elite dairy bulls. As crossbred animals are usually suckler cows or part of the slaughter generation, their current value is not high enough to merit genotyping the large numbers that would be necessary to create a crossbred reference population (it is for this reason that chapter 3 was based on an African dataset as opposed to a UK dataset). Most of the research into crossbred genomic evaluations has been in pigs and poultry due to the structure of their industries (Hidalgo et al., 2015). Because of this, investigation into the utility of using a crossbred reference population for prediction of GEBVs in UK cattle is reliant on research funding that will allow the genotyping of specific crossbred individuals to add to the reference population. One source of these genotypes could be via the Beef Efficiency Scheme (Scottish Government, 2016) which aims to phenotype thousands of crossbred beef calves over the next five years, and genotype

20% of animals recorded. Although this resource will be made up of beef cattle, if research based on this population shows that crossbred genomic evaluations using a crossbred reference population is feasible, then we may see more value placed on the genotyping of crossbred individuals.

5.5 Using sequence data for across breed genomic evaluations

The results of chapter 4 were promising, but to make stronger conclusions regarding the utility of novel SNPs for computing accurate multi-breed genomic evaluations the method would need to be tested in a larger population where there are enough sequenced animals available for a genomic evaluation to be feasible. If such a study were to yield good results, then the next step would be to validate the results in different populations, and a decision would need to be made about whether it would be more appropriate to create custom genotyping chips that directly genotype the SNPs discovered, or whether information on these markers should be determined via imputation from chips to sequence. For either scenario, a large population containing information from multiple breeds would need to be sequenced, not only to determine SNPs of interest across a wider range of cattle breeds, but also to enable precise haplotyping in order to facilitate high accuracy imputation.

At the present time, whole-genome sequencing costs approximately $1000 per individual (M. Watson, Roslin Institute, personal communication) to generate sequence with a read depth of 30x (i.e. each base is read 30 times on average) and so to genotype a sample population of multiple breeds of cattle would still require a large initial outlay. It has been proposed by Hickey (2013) that sequence data could

be generated for a large population of individuals via low coverage sequencing, where a large population of animals could be sequenced at a read depth of 0.1x for less than $30 per animal, which would allow sequencing of more individuals for a fixed overall cost compared to high-coverage sequencing. The idea behind sequencing at low-coverage would be that although the sequence coverage of a particular individual would be low, combining low coverage sequence for millions of individuals would allow the construction of accurate haplotypes which would allow imputation to full sequence for all animals (Hickey, 2013). At the present time, however, reads from current sequencing technologies are not distributed evenly across the genome, which could lead to large gaps in the sequences when sequenced at a low read-depth, even when a large number of animals are sequenced (Hickey, 2013). More research will be necessary and more effective sequencing technologies will need to be developed before this can become a reality. As mentioned in chapter 4, the computational requirements of using WGS data for routine genomic evaluations would mean that unless pertinent data was extracted from sequence in order to compute GEBVs, then novel computing strategies will also need to be developed that can handle the large volumes of data generated by WGS.

Regardless of whether sequencing data is used in its entirety, or simply to allow identification of appropriate SNPs for genomic evaluations, success of multi-breed evaluations will still rely on causal variants having similar effects across the breeds of interest, as discussed previously.

5.6 Availability of genotype data

In all of the analyses carried out in this thesis, the data available to carry out genomic evaluations was limited. This lack of data is likely to be one of the largest obstacles in tackling multi-breed and crossbred genomic evaluations in cattle. A large proportion of published work investigating the potential for genomic selection in multi-breed and crossbred reference populations is based on simulated rather than empirical data. While simulations are an extremely useful tool to predict what will happen when using empirical data, it is not necessarily possible to accurately emulate every property of a real population, and so the results of simulation studies can be somewhat simplistic or unrealistic (Daetwyler et al., 2013). There is a long way to go with regards to effective data collection before multi-breed and crossbred genomic evaluations can be accurately tested and routinely computed for UK cattle populations. Genotyping strategies should be pursued in conjunction with stakeholders such as breed societies, and methods to maximise the value of genotyping such as imputation should be considered, however as discussed previously, care should be taken to ensure high accuracy of imputation.

5.7 Conclusions

Since the introduction of genomic prediction to dairy cattle breeding, the rate of genetic gain has increased for those breeds with sufficient numbers to create single breed reference populations. However, due to differences between breeds, further research has been needed in order to facilitate the commercial implementation of both multi-breed and crossbred genomic evaluations. This thesis has established that both multi-breed and crossbred genomic predictions are feasible using methods

developed from traditional pedigree based evaluations. The methods implemented within this thesis should now be further tested and refined using larger datasets, before implementing across-breed genomic evaluations in a commercial setting.

As the costs associated with genotyping are steadily decreasing, further genotypes may be collected, potentially allowing the development of within breed evaluations for further breeds. However, some breeds will never have sufficient animals to create a single breed reference population; as such, multi-breed reference populations will always be necessary to calculate GEBVs. Methods such as those applied in this thesis will, therefore, be required to exploit these reference populations in order facilitate genetic improvement in numerically small breeds.

References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–752. doi:10.3168/jds.2009-2730.

Bastiaansen, J.W.M., A. Coster, M.P.L. Calus, J.A.M. van Arendonk, and H. Bovenhuis. 2012. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genet. Sel. Evol.* 44:(24 January 2012). doi:Artn 3\rDoi 10.1186/1297-9686-44-3.

Van Den Berg, I., D. Boichard, B. Guldbrandtsen, and M.S. Lund. 2016. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. *G3 Genes | Genomes | Genet.* 6:2553–2561. doi:10.1534/g3.116.027730/-/DC1.

Berry, D.P., M.C. McClure, and M.P. Mullen. 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. *J. Anim. Breed. Genet.* 131:165–172. doi:10.1111/jbg.12067.

van Binsbergen, R., M.P.L. Calus, M.C.A.M. Bink, F.A. van Eeuwijk, C. Schrooten, and R.F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 47:71.

doi:10.1186/s12711-015-0149-x.

Brøndum, R.F., B. Guldbrandtsen, G. Sahana, M.S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics*. 15:728. doi:10.1186/1471-2164-15-728.

Brøndum, R.F., E. Rius-Vilarrasa, I. Strandén, G. Su, B. Guldbrandtsen, W.F. Fikse, and M.S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *J. Dairy Sci.* 94:4700–4707. doi:10.3168/jds.2010-3765.

Calus, M.P.L. 2010. Genomic breeding value prediction: methods and procedures. *Animal*. 4:157. doi:10.1017/S1751731109991352.

Calus, M.P.L., T.H.E. Meuwissen, a. P.W. De Roos, and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*. 178:553–561. doi:10.1534/genetics.107.080838.

Carillier, C., H. Larroque, and C. Robert-Granié. 2014. Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. *Genet. Sel. Evol.* 46:67. doi:10.1186/s12711-014-0067-3.

Christensen, O.F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. *Genet. Sel. Evol.* 46:23. doi:10.1186/1297-9686-46-23.

Cingolani, P. 2012. snpEff Documentation.

Cingolani, P., A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. 2012. A program for annotating and predicting the effects of

single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 6:80–92. doi:10.4161/fly.19695.

Clark, S. a, J.M. Hickey, H.D. Daetwyler, and J.H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44:4. doi:10.1186/1297-9686-44-4.

Cleveland, M.A., J.M. Hickey, and S. Forni. 2013. A Common Dataset for Genomic Analysis of Livestock Populations. *G3*. 2:429–435. doi:10.1534/g3.111.001453.

Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*. 193:347–365. doi:10.1534/genetics.112.147983.

Daetwyler, H.D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R.F. Brøndum, X. Liao, A. Djari, S.C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P.J. Bowman, D. Coote, A.J. Chamberlain, C. Anderson, C.P. VanTassell, I. Hulsegge, M.E. Goddard, B. Guldbrandtsen, M.S. Lund, R.F. Veerkamp, D.A. Boichard, R. Fries, and B.J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Publ. Gr.* 46. doi:10.1038/ng.3034.

Daetwyler, H.D., J.M. Hickey, J.M. Henshall, S. Dominik, B. Gredler, J.H.J. Van Der Werf, and B.J. Hayes. 2010. Accuracy of estimated genomic breeding

values for wool and meat traits in a multi-breed sheep population. *Anim. Prod. Sci.* 50:1004–1010. doi:10.1071/AN10096.

Daetwyler, H.D., K.E. Kemper, J.H.J. van der Werf, and B.J. Hayes. 2012a. Components of the accuracy of genomic selection in a multi-breed sheep population. *J. Anim. Sci.* 25:3375–3384. doi:10.2527/jas2011-4557.

Daetwyler, H.D., A. a Swan, J.H.J. van der Werf, and B.J. Hayes. 2012b. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genet. Sel. Evol.* 44:33. doi:10.1186/1297-9686-44-33.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*. 3. doi:10.1371/journal.pone.0003395.

Dekkers, J.C.M., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. *Nat. Rev. Genet.* 3:22–32. doi:10.1038/nrg701.

Druet, T., I. Macleod, and B. Hayes. 2013. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity (Edinb).* 112:39–47. doi:10.1038/hdy.2013.13.

Druet, T., C. Schrooten, and a P.W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J. Dairy Sci.* 93:5443–5454. doi:10.3168/jds.2010-3255.

England, P.R., J.M. Cornuet, P. Berthier, D.A. Tallmon, and G. Luikart. 2006.

Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv. Genet.* 7:303–308. doi:Doi 10.1007/S10592-005-9103-8.

Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B. a Mason, and M.E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95:4114–29. doi:10.3168/jds.2011-5019.

Ertl, J., C. Edel, R. Emmerling, H. Pausch, R. Fries, and K.-U. Götz. 2014. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: observations from Fleckvieh cattle. *J. Dairy Sci.* 97:487–96. doi:10.3168/jds.2013-6855.

Esfandyari, H., A.C. Sørensen, and P. Bijma. 2015. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.* 47:76. doi:10.1186/s12711-015-0155-z.

Eynard, S.E., J.J. Windig, G. Leroy, R. Van Binsbergen, and M.P. Calus. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. *BMC Genet.* doi:10.1186/s12863-015-0185-0.

Gao, H., O.F. Christensen, P. Madsen, U.S. Nielsen, Y. Zhang, M.S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genet. Sel. Evol.* 44:8. doi:10.1186/1297-9686-44-8.

García-Ruiz, A., J.B. Cole, P.M. VanRaden, G.R. Wiggans, F.J. Ruiz-López, and

C.P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc. Natl. Acad. Sci. United States*. 113:3995–4004.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet. Sel. Evol.* 41:55. doi:10.1186/1297-9686-41-55.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr. arXiv1207.3907*. 9. doi:arXiv:1207.3907 [q-bio.GN].

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. *J. Anim. Breed. Genet.* 124:323–30. doi:10.1111/j.1439-0388.2007.00702.x.

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* 10:381–391. doi:10.1038/nrg2575.

Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. 2010. Genomic selection in livestock populations. *Genet. Res. (Camb).* 92:413–421. doi:10.1017/S0016672310000613.

Goddard, M.E., I.M. Macleod, S. Bolormaa, B.J. Hayes, and K.E. Kemper. 2015. Improving the accuracy of genomic predictions. *In* Proceedings of the Twenty-first Conference of the Association for the Advancement of Animal Breeding and Genetics. 149–152.

Haley, C.S., and P.M. Visscher. 1998. Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci.* 81 Suppl 2:85–97. doi:10.3168/jds.S0022-0302(98)70157-2.

Harris, B.., A.M. Winkelman, and D. Johnston. 2011. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *In* INTERBULL BULLETIN. Stavanger, Norway.

Hay, E.H., and R. Rekaya. 2014. A multi-compartment model for genomic selection in multi-breed populations. *Livest. Sci.* 177:1–7. doi:10.1016/j.livsci.2015.03.027.

Hayes, B., and M.E. Goddard. 2003. Evaluation of marker assisted selection in pig enterprises. *Livest. Prod. Sci.* 81:197–211. doi:10.1016/S0301-6226(02)00257-9.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* 41:51. doi:10.1186/1297-9686-41-51.

Hayes, B.J., P.J. Bowman, a J. Chamberlain, and M.E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92:433–443. doi:10.3168/jds.2008-1646.

Hayes, B.J., I.M. Macleod, H.D. Daetwyler, P.J. Bowman, A.J. Chamberlian, C.J. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M.S. Lund, D.A. Boichard, R.F. Veerkamp, C.P. Vantassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D.J. Dekoning, E. Santus, and M.E. Goddard. 2014. Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. *In* Proceedings of the 10th World Congress in Genetics Applied to Livestock Production. Vancouver.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 31:423–447. doi:10.2307/2529430.

Hickey, J.M. 2013. Sequencing millions of animals for genomic selection 2.0. *J. Anim. Breed. Genet. = Zeitschrift für Tierzüchtung und Züchtungsbiologie*. 130:331–2. doi:10.1111/jbg.12054.

Hidalgo, A.M., J.W.M. Bastiaansen, M.S. Lopes, R. Veroneze, M.A.M. Groenen, and D.J. de Koning. 2015. Accuracy of genomic prediction using deregressed breeding 1 values estimated from purebred and crossbred offspring phenotypes in pigs. *J. Anim. Sci.* 93:3313–3321. doi:10.2527/jas2015-8899.

Hill, W.G. 2010. Understanding and using quantitative genetic variation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365:73–85. doi:10.1098/rstb.2009.0203.

Hill, W.G. 2014. Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics*. 196:1–16. doi:10.1534/genetics.112.147850.

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. *Genet. Sel. Evol.* 45:33. doi:10.1186/1297-9686-45-33.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. *J. Dairy Sci.* in press:1–12. doi:10.3168/jds.2013-7761.

Ibánez-Escriche, N., R.L. Fernando, A. Toosi, and J.C.M. Dekkers. 2009. Genomic

selection of purebreds for crossbred performance. *Genet. Sel. Evol.* 41:12. doi:10.1186/1297-9686-41-12.

Ibáñez-Escriche, N., S. Forni, J.L. Noguera, and L. Varona. 2014. Genomic information in pig breeding: Science meets industry needs. *Livest. Sci.* 166:94–100. doi:10.1016/j.livsci.2014.05.020.

Iheshiulor, O.O.M., J.A. Woolliams, X. Yu, R. Wellmann, and T.H.E. Meuwissen. 2016. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet. Sel. Evol.* 48:15. doi:10.1186/s12711-016-0193-1.

Janss, L., G. de los Campos, N. Sheehan, and D. Sorensen. 2012. Inferences from genomic models in stratified populations. *Genetics*. 192:693–704. doi:10.1534/genetics.112.141143.

Jattawa, D., M. Elzo, S. Koonawootrittriron, and T. Suwanasopee. 2016. Imputation Accuracy from Low to Moderate Density Single Nucleotide Polymorphism Chips in a Thai Multibreed Dairy Cattle Population. *Asian-Australasian J. Anim. Sci.* 29:464–470. doi:10.5713/ajas.15.0291.

Jonas, E., and D.J. de Koning. 2015. Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Front. Genet.* 5. doi:10.3389/fgene.2015.00049.

Karoui, S., M.J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genet. Sel. Evol.* 44:39. doi:10.1186/1297-9686-44-39.

Kizilkaya, K., R.L. Fernando, and D.J. Garrick. 2010. Genomic prediction of

simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88:544–551. doi:10.2527/jas.2009-2064.

Koivula, M., I. Strandén, G. Su, and E. a Mäntysaari. 2012. Different methods to calculate genomic predictions--comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J. Dairy Sci.* 95:4065–4073. doi:10.3168/jds.2011-4874.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. *Livest. Sci.* 166:54–65. doi:10.1016/j.livsci.2014.04.029.

Lidauer, M., and I. Strandén. 1999. Fast and flexible program for genetic evaluation in dairy cattle. *Interbull Bull.* 20:19–24.

Lillehammer, M., T.H.E. Meuwissen, and a K. Sonesson. 2011. A comparison of dairy cattle breeding designs that use genomic selection. *J. Dairy Sci.* 94:493–500. doi:10.3168/jds.2010-3518.

Lorenz, A., K.P. Smith, and A.J. Lorenz. 2015. Adding Genetically Distant Individualsto Training Populations Reduces GenomicPrediction Accuracy in Barley. *Crop Sci.* 55.

Luan, T., J. a. Woolliams, S. Lien, M. Kent, M. Svendsen, and T.H.E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics*. 183:1119–1126. doi:10.1534/genetics.109.107391.

Luan, T., J. a Woolliams, J. Ødegård, M. Dolezal, S.I. Roman-Ponce, A. Bagnato,

and T.H. Meuwissen. 2012. The importance of identity-by-state information for the accuracy of genomic selection. *Genet. Sel. Evol.* 44:28. doi:10.1186/1297-9686-44-28.

Lund, M.S., G. Su, L. Janss, B. Guldbrandtsen, and R.F. Brøndum. 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. *Livest. Sci.* 166:101–110. doi:10.1016/j.livsci.2014.05.008.

Ma, P., M.S. Lund, X. Ding, Q. Zhang, and G. Su. 2014. Increasing imputation and prediction accuracy for Chinese Holsteins using joint Chinese-Nordic reference population. *J. Anim. Breed. Genet.* 131:462–472. doi:10.1111/jbg.12111.

Makgahlela, M.L., E. a. Mäntysaari, I. Strandén, M. Koivula, U.S. Nielsen, M.J. Sillanpää, and J. Juga. 2013. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *J. Anim. Breed. Genet.* 130:10–19. doi:10.1111/j.1439-0388.2012.01017.x.

Matukumalli, L.K., C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P.L. Smith, T.S. Sonstegard, and C.P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS One*. 4. doi:10.1371/journal.pone.0005350.

McKay, S.D., R.D. Schnabel, B.M. Murdoch, L.K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. Dias Neto, C. a Gill, C. Gao, H. Mannen, P. Stothard, Z. Wang, C.P. Van Tassell, J.L. Williams, J.F. Taylor, and S.S. Moore. 2007. Whole genome linkage disequilibrium maps in cattle. *BMC Genet.* 8:74. doi:10.1186/1471-2156-8-74.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for

complex traits by whole genome resequencing. *Genetics*. 185:623–631. doi:10.1534/genetics.110.116590.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157:1819–1829. doi:11290733.

Mingay, G. 1982. British Friesians: An Epic of Progress. 1st ed. British Friesian Society of Great Britain and Ireland, Rickmansworth, UK.

Misztal, I. 2011. FAQ for genomic selection. *J. Anim. Breed. Genet.* 128:245–246. doi:10.1111/j.1439-0388.2011.00944.x.

Misztal, I., a Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92:4648–4655. doi:10.3168/jds.2009-2064.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. Lee. 2002. BLUPF90 and related programs (BGF90). *In* Proceedings of the 7th world congress on genetics applied to livestock production. Montpellier. 21–22.

Mrode, R.. 2014. Linear Models for the Prediction of Animal Breeding Values. 3rd ed. CABI. 199-200 pp.

Mucha, S., R. Mrode, I. MacLaren-Lee, M. Coffey, and J. Conington. 2015. Estimation of genomic breeding values for milk yield in UK dairy goats. *J. Dairy Sci.* 98:8201–8208. doi:10.3168/jds.2015-9682.

Mujibi, F.D.N., J.D. Nkrumah, O.N. Durunna, P. Stothard, J. Mah, Z. Wang, J. Basarab, G. Plastow, D.H. Crews, and S.S. Moore. 2011. Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. *J. Anim. Sci.*

89:3353–3361. doi:10.2527/jas.2010-3361.

Nadaf, J., V. Riggio, T.-P. Yu, and R. Pong-Wong. 2012. Effect of the prior distribution of SNP effects on the estimation of total breeding value. *BMC Proc.* 6:S6. doi:10.1186/1753-6561-6-S2-S6.

Ojango, J.M.K., A. Marete, D. Mujibi, J. Rao, J. Pool, J.E.O. Rege, C. Gondro, W.M.S.P. Weerasinghe, J.P. Gibson, and A.M. Okeyo. 2014. A novel use of high density SNP assays to optimize choice of different crossbred dairy cattle genotypes in small-holder systems in East Africa. *Proc. 10th World Congr. Genet. Appl. to Livest. Prod.* 2–4.

Olson, K.M., P.M. VanRaden, and M.E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.* 95:5378–83. doi:10.3168/jds.2011-5006.

Orel, V. 1997. Cloning, inbreeding, and history. *Q. Rev. Biol.* 72:437–40.

Pimentel, E.C.G., C. Edel, R. Emmerling, and K.-U. Götz. 2015. How imputation errors bias genomic predictions. *J. Dairy Sci.* 98:4131–8. doi:10.3168/jds.2014-9170.

Pryce, J.E., B. Gredler, S. Bolormaa, P.J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M.E. Goddard, and B.J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. *J. Dairy Sci.* 94:2625–2630. doi:10.3168/jds.2010-3719.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses.

Am. J. Hum. Genet. 81:559–575. doi:10.1086/519795.

R Core Team. 2013. R Core Team. *R A Lang. Environ. Stat. Comput. R Found. Stat. Comput. Vienna, Austria.* ISBN 3-900051-07-0, URL http://www.R-project.org/.

Riedelsheimer, C., J.B. Endelman, M. Stange, M.E. Sorrells, J.-L. Jannink, and A.E. Melchinger. 2013. Genomic Predictability of Interconnected Biparental Maize Populations. *Genetics*. 194:493–503. doi:10.1534/genetics.113.150227.

Rodríguez-Ramilo, S.T., L.A. García-Cortés, and M.Á.R. de Cara. 2015. Artificial selection with traditional or genomic relationships: consequences in coancestry and genetic diversity. *Front. Genet.* 6:127. doi:10.3389/fgene.2015.00127.

Rolf, M.M., J.F. Taylor, R.D. Schnabel, S.D. Mckay, M.C. Mcclure, S.L. Northcutt, M.S. Kerley, and R.L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genet.* 11.

De Roos, A.P.W., B.J. Hayes, R.J. Spelman, and M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*. 179:1503–1512. doi:10.1534/genetics.107.084301.

De Roos, a. P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics*. 183:1545–1553. doi:10.1534/genetics.109.104935.

de Roos, a P.W., C. Schrooten, and T. Druet. 2011. Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. *J. Dairy Sci.* 94:4708–4714. doi:10.3168/jds.2010-

3905.

Sánchez-Molano, E., D. Tsiokos, D. Chatziplis, H. Jorjani, L. Degano, C. Diaz, A. Rossoni, H. Schwarzenbacher, F. Seefried, L. Varona, D. Vicario, E.L. Nicolazzi, and G. Banos. 2016. A practical approach to detect ancestral haplotypes in livestock populations. *BMC Genet.* 17:91. doi:10.1186/s12863-016-0405-2.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123:218–223. doi:10.1111/j.1439-0388.2006.00595.x.

Scotti, E., L. Fontanesi, F. Schiavini, V. La Mattina, A. Bagnato, and V. Russo. 2010. DGAT1 p.K232A polymorphism in dairy and dual purpose Italian cattle breeds. *Ital. J. Anim. Sci.* 9.

Scottish Government. 2016. Beef Efficiency Scheme full guidance.

Simeone, R., I. Misztal, I. Aguilar, and a. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. *J. Anim. Breed. Genet.* 128:386–393. doi:10.1111/j.1439-0388.2011.00926.x.

Sonesson, A.K., J. a Woolliams, and T.H. Meuwissen. 2012. Genomic selection requires genomic control of inbreeding. *Genet. Sel. Evol.* 44:27. doi:10.1186/1297-9686-44-27.

Su, G., R.F. Brøndum, P. Ma, B. Guldbrandtsen, G.P. Aamand, and M.S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *J. Dairy Sci.* 95:4657–65.

doi:10.3168/jds.2012-5379.

Suchocki, T., A. Żarnecki, and J. Szyda. 2014. Do rare variants contribute to the genomic prediction accuracy? *In* Proceedings of 10th World Congress of Genetics Applied to Livestock Production. Vancouver.

Taylor, J.F., K.H. Taylor, and J.E. Decker. 2016. Holsteins are the genomic selection poster cows. *Proc. Natl. Acad. Sci. United States*. 113:7690–7692. doi:10.1073/pnas.1608144113.

Thomasen, J.R., C. Egger-Danner, a Willam, B. Guldbrandtsen, M.S. Lund, and a C. Sørensen. 2014. Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. *J. Dairy Sci.* 97:458–70. doi:10.3168/jds.2013-6599.

Tier, B., K. Meyer, and M.H. Ferdosi. 2015. Which Genomic Relationship Matrix? *In* Proceedings of the Twenty-first Conference of the Association for the Advancement of Animal Breeding and Genetics. Lorne, Australia. 461–464.

Toosi, a., R.L. Fernando, and J.C.M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. *J. Anim. Sci.* 88:32–46. doi:10.2527/jas.2009-1975.

Vallée, a, J. a M. van Arendonk, and H. Bovenhuis. 2014. Accuracy of genomic prediction when combining two related crossbred populations. *J. Anim. Sci.* 4342–4348. doi:10.2527/jas.2014-8109.

VanDoorMal, B. 2009. The Road to Genomic Evaluations in Canada.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980.

Vanraden, P.M., and T.A. Cooper. 2015. Genomic Evaluations and Breed

Composition for Crossbred U . S . Dairy Cattle. 9–13.

VanRaden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans, M.E. Tooker, J.B. Cole, T.S. Sonstegard, E.E. Connor, M. Winters, J.B.C.H.M. van Kaam, a Valentini, B.J. Van Doormaal, M. a Faust, and G. a Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* 96:668–78. doi:10.3168/jds.2012-5702.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. doi:10.3168/jds.2008-1514.

Vickers, M., C. Brown, and L. Ford. 2014. Beef production from the dairy herd. Huntingdon.

Villumsen, T.M., L. Janss, and M.S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126:3–13. doi:10.1111/j.1439-0388.2008.00747.x.

Wang, C., D. Habier, B.L. Peiris, a Wolc, a Kranis, K. a Watson, S. Avendano, D.J. Garrick, R.L. Fernando, S.J. Lamont, and J.C.M. Dekkers. 2013. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. *Poult. Sci.* 92:1712–23. doi:10.3382/ps.2012-02941.

Waples, R.S., and C. Do. 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. *Evol. Appl.* 3:244–262. doi:10.1111/j.1752-

4571.2009.00104.x.

Weber, K.L., R.M. Thallman, J.W. Keele, W.M. Snelling, G.L. Bennett, T.P.L. Smith, T.G. McDaneld, M.F. Allan, a. L. Van Eenennaam, and L. a. Kuehn. 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. *J. Anim. Sci.* 90:4177–4190. doi:10.2527/jas.2011-4586.

Weng, Z., Z. Zhang, Q. Zhang, W. Fu, S. He, and X. Ding. 2013. Comparison of different imputation methods from low- to high-density panels using Chinese Holstein cattle. *Animal*. 7:729–735. doi:10.1017/S1751731112002224.

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*. 193:621–631. doi:10.1534/genetics.112.146290.

Wolc, A., J. Arango, P. Settar, J.E. Fulton, N.P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D.J. Garrick, and J.C.M. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet. Sel. Evol.* 43:23. doi:10.1186/1297-9686-43-23.

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O.F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. *J. Anim. Sci.* 94:936. doi:10.2527/jas.2015-9930.

Zeng, J., A. Toosi, R.L. Fernando, J.C.M. Dekkers, and D.J. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.* 45:11. doi:10.1186/1297-9686-45-11.

Zenger, K.R., M.S. Khatkar, J.A.L. Cavanagh, R.J. Hawken, and H.W. Raadsma.

2007. Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. *Anim. Genet.* 38:7–14. doi:10.1111/j.1365-2052.2006.01543.x.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M.S. Lund, and G. Su. 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. *Genet. Sel. Evol.* 45:7. doi:10.1186/1297-9686-45-7.

Zhou, L., B. Heringstad, G. Su, B. Guldbrandtsen, T.H.E. Meuwissen, M. Svendsen, H. Grove, U.S. Nielsen, and M.S. Lund. 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *J. Dairy Sci.* 97:4485–96. doi:10.3168/jds.2013-7580.

Zhou, L., M.S. Lund, Y. Wang, and G. Su. 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *J. Anim. Breed. Genet.* 131:249–257. doi:10.1111/jbg.12089.